



June 2003, cloth,
0-7167-9623-6

[Companion Site](#)

- ▶ [Summary](#)
- ▶ [Features](#)
- ▶ [New to This Edition](#)
- ▶ [Media](#)
- ▶ [Supplements](#)
- ▶ [Table of Contents](#)
- ▶ [Preview Materials](#)

▶ Other Titles by:
▶ [David S. Moore](#)

The Basic Practice of Statistics

Third Edition

David S. Moore (Purdue U.)

NEW ▶ [Preview this Book](#)

▶ [Request Exam Copy](#)

▶ [Go To Companion Site](#)

Download Text chapters in .PDF format.



You will need Adobe Acrobat Reader version 3.0 or above to view these preview materials. (**Additional instructions below.**)

Exploring Data: Variables and Distributions

[Chapter 1 - Picturing Distributions with Graphs](#) (CH 01.pdf; 300KB)

[Chapter 2 - Describing Distributions with Numbers](#) (CH 02.pdf; 212KB)

[Chapter 3 - Normal Distributions](#) (CH 03.pdf; 328KB)

Exploring Data: Relationships

[Chapter 4 - Scatterplots and Correlation](#) (CH 04.pdf; 300KB)

[Chapter 5 - Regression](#) (CH 05.pdf; 212KB)

[Chapter 6 - Two-Way Tables](#) (CH 06.pdf; 328KB)

These copyrighted materials are for promotional purposes only. They may not be sold, copied, or distributed.

Download Instructions for Preview Materials in .PDF Format

We recommend saving these files to your hard drive by following the instructions below.

PC users

1. Right-click on a chapter link below
2. From the pop-up menu, select "Save Link", (if you are using Netscape) or "Save Target" (if you are using Internet Explorer)
3. In the "Save As" dialog box, select a location on your hard drive and rename the file, if you would like, then click "save". Note the name and location of the file so you can open it later.

Macintosh users

1. Click and hold your mouse on a chapter link below
2. From the pop-up menu, select "Save Link As" (if you are using Netscape) or "Save Target As" (if you are using Internet Explorer)
3. In the "Save As" dialog box, select a location on your hard drive and rename the file, if you would like, then click "save". Note the name and location of the file so you can open it later.

VISIT...

LANZAROTE
Caliente.COM

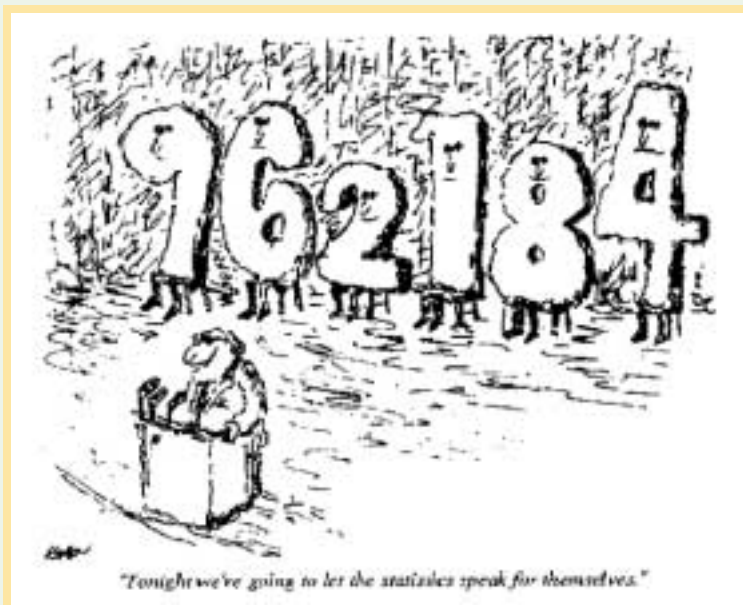
Exploring Data

The first step in understanding data is to hear what the data say, to “let the statistics speak for themselves.” But numbers speak clearly only when we help them speak by organizing, displaying, summarizing, and asking questions. That’s *data analysis*. The six chapters in Part I present the ideas and tools of statistical data analysis. They equip you with skills that are immediately useful whenever you deal with numbers.

These chapters reflect the strong emphasis on exploring data that characterizes modern statistics. Although careful exploration of data is essential if we are to trust the results of inference, data analysis isn’t just preparation for inference. To think about inference, we carefully distinguish between the data we actually have and the larger universe we want conclusions about. The Bureau of Labor Statistics, for example, has data about employment in the 55,000 households contacted by its Current Population Survey. The bureau wants to draw conclusions about employment in all 110 million U.S. households. That’s a complex problem. From the viewpoint of data analysis, things are simpler. We want to explore and understand only the data in hand. The distinctions that inference requires don’t concern us in Chapters 1 to 6. What does concern us is a systematic strategy for examining data and the tools that we use to carry out that strategy.

Part of that strategy is to first look at one thing at a time and then at relationships. In Chapters 1, 2, and 3 you will study **variables and their distributions**. Chapters 4, 5, and 6 concern **relationships among variables**.

PART



EXPLORING DATA: VARIABLES AND DISTRIBUTIONS

Chapter 1 Picturing Distributions with Graphs

Chapter 2 Describing Distributions with Numbers

Chapter 3 The Normal Distributions

EXPLORING DATA: RELATIONSHIPS

Chapter 4 Scatterplots and Correlation

Chapter 5 Regression

Chapter 6 Two-Way Tables

EXPLORING DATA REVIEW

CHAPTER

1

(Darrell Ingham/Allsport Concepts/Getty Images)



Picturing Distributions with Graphs

Statistics is the science of data. The volume of data available to us is overwhelming. Each March, for example, the Census Bureau collects economic and employment data from more than 200,000 people. From the bureau's Web site you can choose to examine more than 300 items of data for each person (and more for households): child care assistance, child care support, hours worked, usual weekly earnings, and much more. The first step in dealing with such a flood of data is to organize our thinking about data.

Individuals and variables

Any set of data contains information about some group of *individuals*. The information is organized in *variables*.

INDIVIDUALS AND VARIABLES

Individuals are the objects described by a set of data. Individuals may be people, but they may also be animals or things.

A **variable** is any characteristic of an individual. A variable can take different values for different individuals.

In this chapter we cover...

Individuals and variables

Categorical variables:
pie charts and bar graphs

Quantitative variables:
histograms

Interpreting histograms

Quantitative variables:
stemplots

Time plots



Are data artistic?

David Galenson, an economist at the University of Chicago, uses data and statistical analysis to study innovation among painters from the nineteenth century to the present. Economics journals publish his work. Art history journals send it back unread. “Fundamentally antagonistic to the way humanists do their work,” said the chair of art history at Chicago. If you are a student of the humanities, reading this statistics text may help you start a new wave in your field.

A college’s student data base, for example, includes data about every currently enrolled student. The students are the individuals described by the data set. For each individual, the data contain the values of variables such as date of birth, gender (female or male), choice of major, and grade point average. In practice, any set of data is accompanied by background information that helps us understand the data. When you plan a statistical study or explore data from someone else’s work, ask yourself the following questions:

1. **Who?** What **individuals** do the data describe? **How many** individuals appear in the data?
2. **What?** How many **variables** do the data contain? What are the **exact definitions** of these variables? In what **units of measurement** is each variable recorded? Weights, for example, might be recorded in pounds, in thousands of pounds, or in kilograms.
3. **Why? What purpose** do the data have? Do we hope to answer some specific questions? Do we want to draw conclusions about individuals other than the ones we actually have data for? Are the variables suitable for the intended purpose?

Some variables, like gender and college major, simply place individuals into categories. Others, like height and grade point average, take numerical values for which we can do arithmetic. It makes sense to give an average income for a company’s employees, but it does not make sense to give an “average” gender. We can, however, count the numbers of female and male employees and do arithmetic with these counts.

CATEGORICAL AND QUANTITATIVE VARIABLES

A **categorical variable** places an individual into one of several groups or categories.

A **quantitative variable** takes numerical values for which arithmetic operations such as adding and averaging make sense.

The **distribution** of a variable tells us what values it takes and how often it takes these values.

EXAMPLE 1.1 A professor’s data set

Here is part of the data set in which a professor records information about student performance in a course:

	A	B	C	D	E	F	G	H	I
1	Name	School	Major	HW total	Midterm	Final Exam	Total	Grade	
2	Ansys, Juan	EDU	EdPsych	95	80	88	263	A	
3	Admir, Brenda	LA	Psych	32	61	54	147	D	
4	Bai, Jingqi	SCI	Biol	74	68	70	212	B	
5	Beggs, Amanda	SCI	Math	86	75	94	255	A	

The *individuals* described are the students. Each row records data on one individual. Each column contains the values of one *variable* for all the individuals. In addition to the student’s name, there are 7 variables. School and major are categorical variables. Scores on homework, the midterm, and the final exam and the total score are quantitative. Grade is recorded as a category (A, B, and so on), but each grade also corresponds to a quantitative score ($A = 4$, $B = 3$, and so on) that is used to calculate student grade point averages.

Most data tables follow this format—each row is an individual, and each column is a variable. This data set appears in a **spreadsheet** program that has rows and columns ready for your use. Spreadsheets are commonly used to enter and transmit data and to do simple calculations such as adding homework, midterm, and final scores to get total points.

spreadsheet

APPLY YOUR KNOWLEDGE

1.1 **Fuel economy.** Here is a small part of a data set that describes the fuel economy (in miles per gallon) of 2002 model motor vehicles:

Make and model	Vehicle type	Transmission type	Number of cylinders	City MPG	Highway MPG
⋮					
Acura NSX	Two-seater	Automatic	6	17	24
Audi A4	Compact	Manual	4	22	31
Buick Century	Midsized	Automatic	6	20	29
Dodge Ram 1500	Standard pickup truck	Automatic	8	15	20
⋮					

- (a) What are the individuals in this data set?
- (b) For each individual, what variables are given? Which of these variables are categorical and which are quantitative?

1.2 **A medical study.** Data from a medical study contain values of many variables for each of the people who were the subjects of the study. Which of the following variables are categorical and which are quantitative?

- (a) Gender (female or male)
- (b) Age (years)
- (c) Race (Asian, black, white, or other)
- (d) Smoker (yes or no)
- (e) Systolic blood pressure (millimeters of mercury)
- (f) Level of calcium in the blood (micrograms per milliliter)

exploratory data analysis

Categorical variables: pie charts and bar graphs

Statistical tools and ideas help us examine data in order to describe their main features. This examination is called **exploratory data analysis**. Like an explorer crossing unknown lands, we want first to simply describe what we see. Here are two basic strategies that help us organize our exploration of a set of data:

- Begin by examining each variable by itself. Then move on to study the relationships among the variables.
- Begin with a graph or graphs. Then add numerical summaries of specific aspects of the data.

We will follow these principles in organizing our learning. Chapters 1 to 3 present methods for describing a single variable. We study relationships among several variables in Chapters 4 to 6. In each case, we begin with graphical displays, then add numerical summaries for more complete description.

The proper choice of graph depends on the nature of the variable. The values of a categorical variable are labels for the categories, such as “male” and “female.” The distribution of a categorical variable lists the categories and gives either the **count** or the **percent** of individuals who fall in each category.

EXAMPLE 1.2 Garbage

The formal name for garbage is “municipal solid waste.” Here is a breakdown of the materials that made up American municipal solid waste in 2000.¹

Material	Weight (million tons)	Percent of total
Food scraps	25.9	11.2%
Glass	12.8	5.5%
Metals	18.0	7.8%
Paper, paperboard	86.7	37.4%
Plastics	24.7	10.7%
Rubber, leather, textiles	15.8	6.8%
Wood	12.7	5.5%
Yard trimmings	27.7	11.9%
Other	7.5	3.2%
Total	231.9	100.0

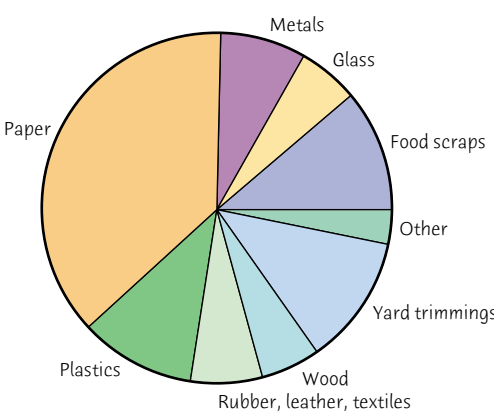
roundoff error

It’s a good idea to check data for consistency. The weights of the nine materials add to 231.8 million tons, not exactly equal to the total of 231.9 million tons given in the table. What happened? **Roundoff error:** Each entry is rounded to the nearest tenth, and the total is rounded separately. The exact values would add exactly, but the rounded values don’t quite.

pie chart

The **pie chart** in Figure 1.1 shows us each material as a part of the whole. For example, the “plastics” slice makes up 10.7% of the pie because 10.7% of municipal solid waste consists of plastics. The graph shows more clearly than the numbers the predominance of paper and the importance of food scraps,

Figure 1.1 Pie chart of materials in municipal solid waste, by weight.



plastics, and yard trimmings in our garbage. Pie charts are awkward to make by hand, but software will do the job for you.

We could also make a **bar graph** that represents each material’s weight by the height of a bar. To make a pie chart, you must include all the categories that make up a whole. Bar graphs are more flexible. Figure 1.2(a) is a bar graph of the percent of each material that was recycled or composted in 2000. These percents are not part of a whole because each refers to a different material. We could replace the pie chart in Figure 1.1 by a bar graph, but we can’t make a pie chart to replace Figure 1.2(a). We can often improve a bar graph by changing the order of the groups we are comparing. Figure 1.2(b) displays the recycling data with the materials in order of percent recycled or composted. Figures 1.1 and 1.2 together suggest that we might pay more attention to recycling plastics.

bar graph

Bar graphs and pie charts help an audience grasp the distribution quickly. They are, however, of limited use for data analysis because it is easy to understand data on a single categorical variable without a graph. We will move on to quantitative variables, where graphs are essential tools.

APPLY YOUR KNOWLEDGE

1.3 The color of your car. Here is a breakdown of the most popular colors for vehicles made in North America during the 2001 model year:²

Color	Percent	Color	Percent
Silver	21.0%	Medium red	6.9%
White	15.6%	Brown	5.6%
Black	11.2%	Gold	4.5%
Blue	9.9%	Bright red	4.3%
Green	7.6%	Grey	2.0%

- (a) What percent of vehicles are some other color?
- (b) Make a bar graph of the color data. Would it be correct to make a pie chart if you added an “Other” category?

The height of this bar is 45.4 because 45.4% of paper municipal waste was recycled.

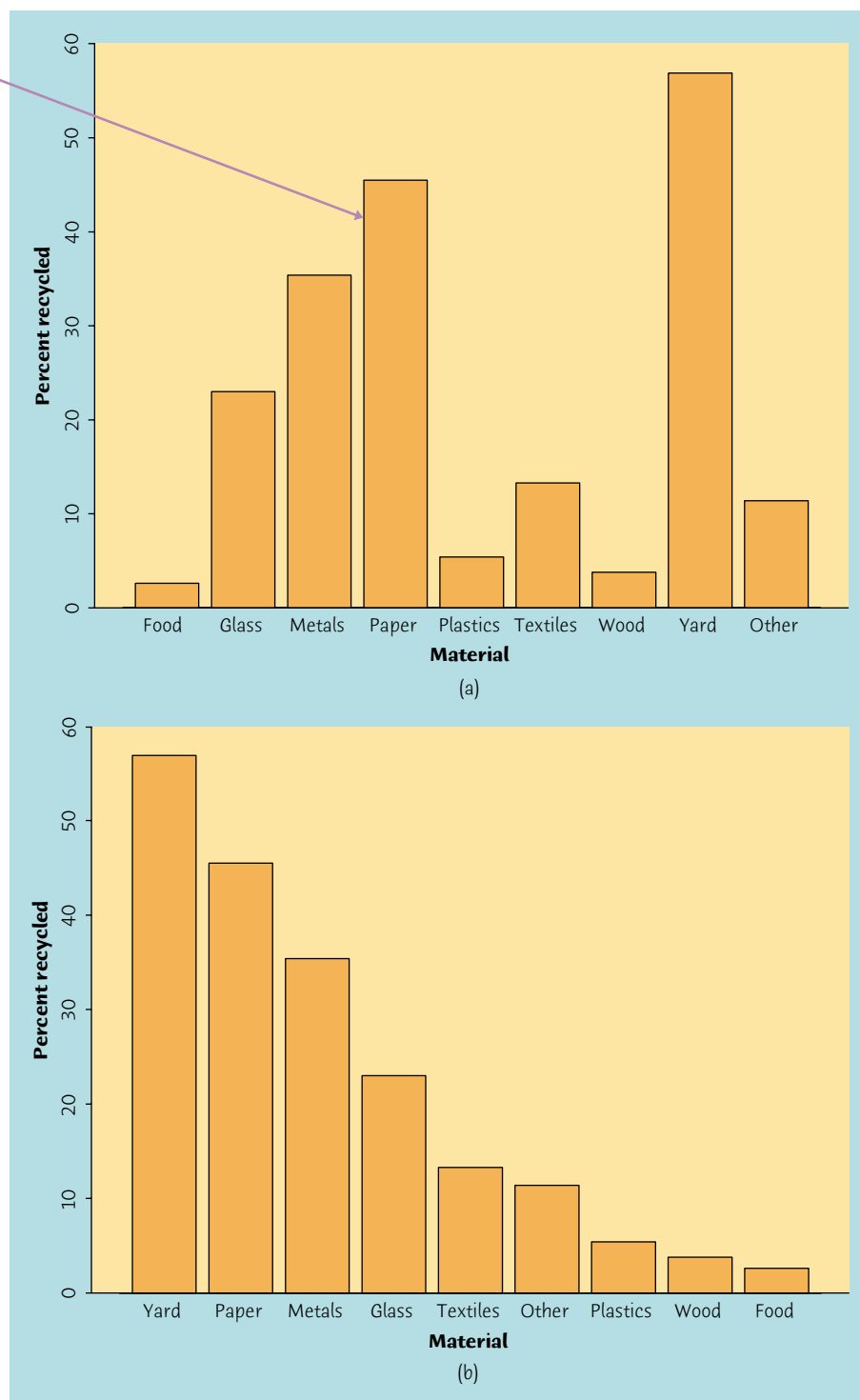


Figure 1.2 Bar graphs comparing the percents of each material in municipal solid waste that were recycled or composted.

1.4 **Never on Sunday?** Births are not, as you might think, evenly distributed across the days of the week. Here are the average numbers of babies born on each day of the week in 1999:³

Day	Births
Sunday	7,731
Monday	11,018
Tuesday	12,424
Wednesday	12,183
Thursday	11,893
Friday	12,012
Saturday	8,654

Present these data in a well-labeled bar graph. Would it also be correct to make a pie chart? Suggest some possible reasons why there are fewer births on weekends.

Quantitative variables: histograms

Quantitative variables often take many values. A graph of the distribution is clearer if nearby values are grouped together. The most common graph of the distribution of one quantitative variable is a **histogram**.

histogram

EXAMPLE 1.3 Making a histogram

One of the most striking findings of the 2000 census was the growth of the Hispanic population of the United States. Table 1.1 presents the percent of residents in each of the 50 states who identified themselves in the 2000 census as “Spanish/Hispanic/Latino.”⁴ The *individuals* in this data set are the 50 states. The *variable* is the percent of Hispanics in a state’s population. To make a histogram of the distribution of this variable, proceed as follows:

Step 1. Choose the classes. Divide the range of the data into classes of equal width. The data in Table 1.1 range from 0.7 to 42.1, so we decide to choose these classes:

0.0 ≤ percent Hispanic < 5.0
5.0 ≤ percent Hispanic < 10.0
⋮
40.0 ≤ percent Hispanic < 45.0

Be sure to specify the classes precisely so that each individual falls into exactly one class. A state with 4.9% Hispanic residents would fall into the first class, but a state with 5.0% falls into the second.

TABLE 1.1 Percent of population of Hispanic origin, by state (2000)

State	Percent	State	Percent	State	Percent
Alabama	1.5	Louisiana	2.4	Ohio	1.9
Alaska	4.1	Maine	0.7	Oklahoma	5.2
Arizona	25.3	Maryland	4.3	Oregon	8.0
Arkansas	2.8	Massachusetts	6.8	Pennsylvania	3.2
California	32.4	Michigan	3.3	Rhode Island	8.7
Colorado	17.1	Minnesota	2.9	South Carolina	2.4
Connecticut	9.4	Mississippi	1.3	South Dakota	1.4
Delaware	4.8	Missouri	2.1	Tennessee	2.0
Florida	16.8	Montana	2.0	Texas	32.0
Georgia	5.3	Nebraska	5.5	Utah	9.0
Hawaii	7.2	Nevada	19.7	Vermont	0.9
Idaho	7.9	New Hampshire	1.7	Virginia	4.7
Illinois	10.7	New Jersey	13.3	Washington	7.2
Indiana	3.5	New Mexico	42.1	West Virginia	0.7
Iowa	2.8	New York	15.1	Wisconsin	3.6
Kansas	7.0	North Carolina	4.7	Wyoming	6.4
Kentucky	1.5	North Dakota	1.2		

Step 2. Count the individuals in each class. Here are the counts:

Class	Count	Class	Count	Class	Count
0.0 to 4.9	27	15.0 to 19.9	4	30.0 to 34.9	2
5.0 to 9.9	13	20.0 to 24.9	0	35.0 to 39.9	0
10.0 to 14.9	2	25.0 to 29.9	1	40.0 to 44.9	1

Step 3. Draw the histogram. Mark the scale for the variable whose distribution you are displaying on the horizontal axis. That’s the percent of a state’s population who are Hispanic. The scale runs from 0 to 45 because that is the span of the classes we chose. The vertical axis contains the scale of counts. Each bar represents a class. The base of the bar covers the class, and the bar height is the class count. There is no horizontal space between the bars unless a class is empty, so that its bar has height zero. Figure 1.3 is our histogram.

The bars of a histogram should cover the entire range of values of a variable. When the possible values of a variable have gaps between them, extend the bases of the bars to meet halfway between two adjacent possible values. For example, in a histogram of the ages in years of university faculty, the bars representing 25 to 29 years and 30 to 34 years would meet at 29.5.

Our eyes respond to the *area* of the bars in a histogram.⁵ Because the classes are all the same width, area is determined by height and all classes are fairly represented. There is no one right choice of the classes in a histogram. Too

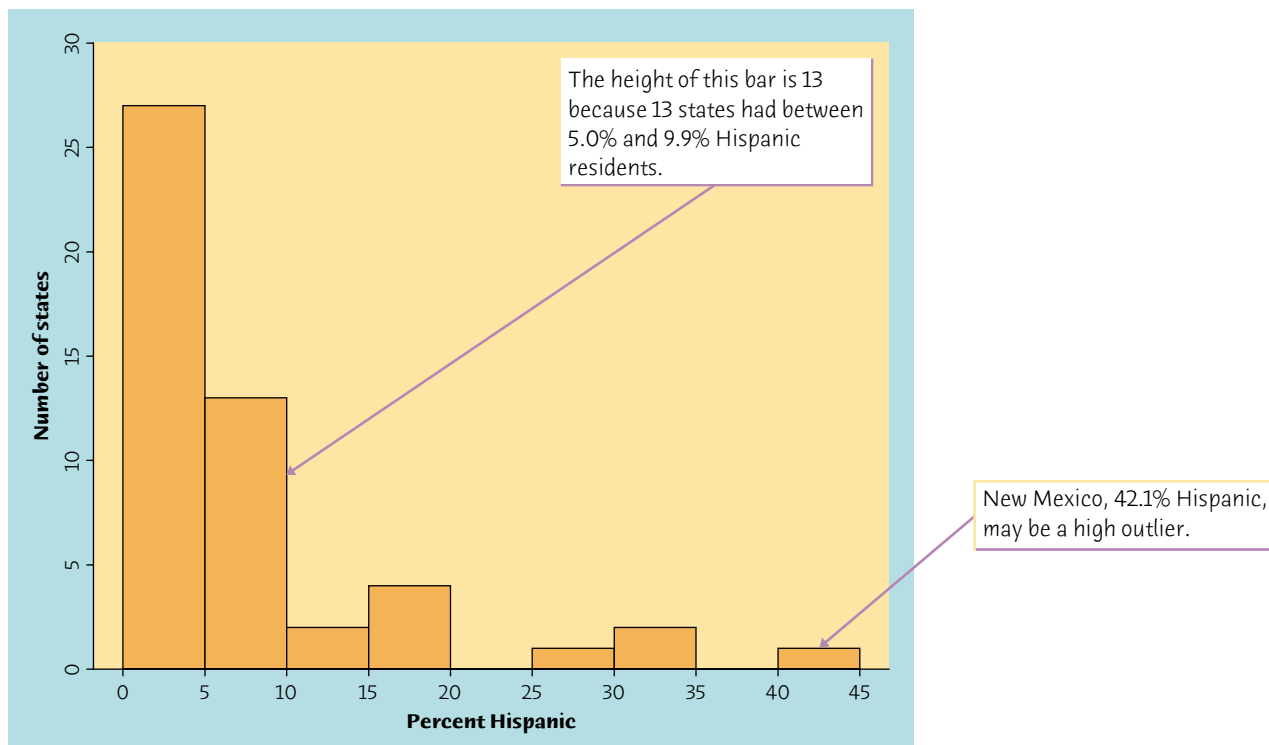


Figure 1.3 Histogram of the distribution of the percent of Hispanics among the residents of the 50 states. This distribution is skewed to the right.

few classes will give a “skyscraper” graph, with all values in a few classes with tall bars. Too many will produce a “pancake” graph, with most classes having one or no observations. Neither choice will give a good picture of the shape of the distribution. You must use your judgment in choosing classes to display the shape. Statistics software will choose the classes for you. The software’s choice is usually a good one, but you can change it if you want.

APPLY YOUR KNOWLEDGE

1.5 Sports car fuel economy. Interested in a sports car? The Environmental Protection Agency lists most such vehicles in its “two-seater” category. Table 1.2 gives the city and highway mileages (miles per gallon) for the 22 two-seaters listed for the 2002 model year.⁶ Make a histogram of the highway mileages of these cars using classes with width 5 miles per gallon.

Interpreting histograms

Making a statistical graph is not an end in itself. The purpose of the graph is to help us understand the data. After you make a graph, always ask, “What do I see?” Once you have displayed a distribution, you can see its important features as follows.

TABLE 1.2 Gas mileage (miles per gallon) for 2002 model two-seater cars

Model	City	Highway	Model	City	Highway
Acura NSX	17	24	Honda Insight	57	56
Audi TT Quattro	20	28	Honda S2000	20	26
Audi TT Roadster	22	31	Lamborghini Murcielago	9	13
BMW M Coupe	17	25	Mazda Miata	22	28
BMW Z3 Coupe	19	27	Mercedes-Benz SL500	16	23
BMW Z3 Roadster	20	27	Mercedes-Benz SL600	13	19
BMW Z8	13	21	Mercedes-Benz SLK230	23	30
Chevrolet Corvette	18	25	Mercedes-Benz SLK320	20	26
Chrysler Prowler	18	23	Porsche 911 GT2	15	22
Ferrari 360 Modena	11	16	Porsche Boxster	19	27
Ford Thunderbird	17	23	Toyota MR2	25	30

EXAMINING A DISTRIBUTION

In any graph of data, look for the **overall pattern** and for striking **deviations** from that pattern.

You can describe the overall pattern of a histogram by its **shape**, **center**, and **spread**.

An important kind of deviation is an **outlier**, an individual value that falls outside the overall pattern.

We will learn how to describe center and spread numerically in Chapter 2. For now, we can describe the center of a distribution by its *midpoint*, the value with roughly half the observations taking smaller values and half taking larger values. We can describe the spread of a distribution by giving the *smallest and largest values*.

EXAMPLE 1.4 Describing a distribution

Look again at the histogram in Figure 1.3. **Shape:** The distribution has a *single peak*, which represents states that are less than 5% Hispanic. The distribution is *skewed to the right*. Most states have no more than 10% Hispanics, but some states have much higher percentages, so that the graph trails off to the right. **Center:** Table 1.1 shows that about half the states have less than 4.7% Hispanics among their residents and half have more. So the midpoint of the distribution is close to 4.7%. **Spread:** The spread is from about 0% to 42%, but only four states fall above 20%.

Outliers: Arizona, California, New Mexico, and Texas stand out. Whether these are outliers or just part of the long right tail of the distribution is a matter of judgment. There is no rule for calling an observation an outlier. Once you have spotted possible outliers, look for an explanation. Some outliers are due to mistakes, such as typing 4.2 as 42. Other outliers point to the special nature of some observations. These four states are heavily Hispanic by history and location.

When you describe a distribution, concentrate on the main features. Look for major peaks, not for minor ups and downs in the bars of the histogram. Look for clear outliers, not just for the smallest and largest observations. Look for rough *symmetry* or clear *skewness*.

SYMMETRIC AND SKEWED DISTRIBUTIONS

A distribution is **symmetric** if the right and left sides of the histogram are approximately mirror images of each other.

A distribution is **skewed to the right** if the right side of the histogram (containing the half of the observations with larger values) extends much farther out than the left side. It is **skewed to the left** if the left side of the histogram extends much farther out than the right side.

Here are more examples of describing the overall pattern of a histogram.

EXAMPLE 1.5 Iowa Test scores

Figure 1.4 displays the scores of all 947 seventh-grade students in the public schools of Gary, Indiana, on the vocabulary part of the Iowa Test of Basic Skills. The

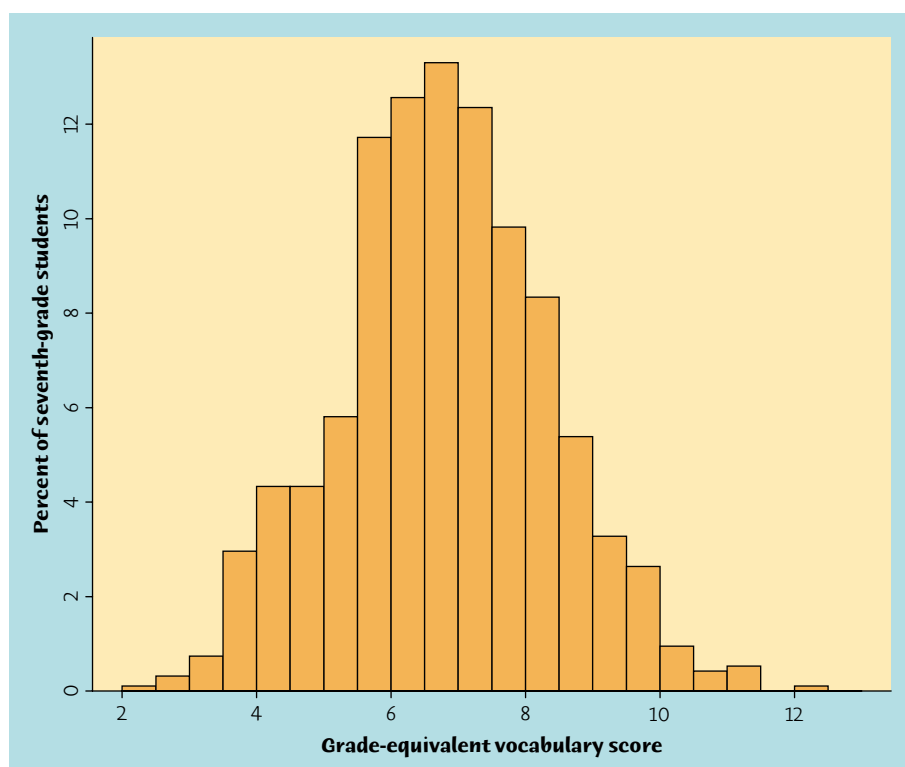


Figure 1.4 Histogram of the Iowa Test vocabulary scores of all seventh-grade students in Gary, Indiana. This distribution is single-peaked and symmetric.

distribution is *single-peaked* and *symmetric*. In mathematics, the two sides of symmetric patterns are exact mirror images. Real data are almost never exactly symmetric. We are content to describe Figure 1.4 as symmetric. The center (half above, half below) is close to 7. This is seventh-grade reading level. The scores range from 2.0 (second-grade level) to 12.1 (twelfth-grade level).

Notice that the vertical scale in Figure 1.4 is not the *count* of students but the *percent* of Gary students in each histogram class. A histogram of percents rather than counts is convenient when we want to compare several distributions. To compare Gary with Los Angeles, a much bigger city, we would use percents so that both histograms have the same vertical scale.

EXAMPLE 1.6 College costs

clusters

Jeanna plans to attend college in her home state of Massachusetts. In the College Board's *Annual Survey of Colleges*, she finds data on estimated college costs for the 2002–2003 academic year. Figure 1.5 displays the costs for all 56 four-year colleges in Massachusetts (omitting art schools and other special colleges). As is often the case, we can't call this irregular distribution either symmetric or skewed. The big feature of the overall pattern is two separate **clusters** of colleges, 11 costing less than \$16,000 and the remaining 45 costing more than \$20,000. Clusters suggest that two types of individuals are mixed in the data set. In fact, the histogram distinguishes the 11 state colleges in Massachusetts from the 45 private colleges, which charge much more.

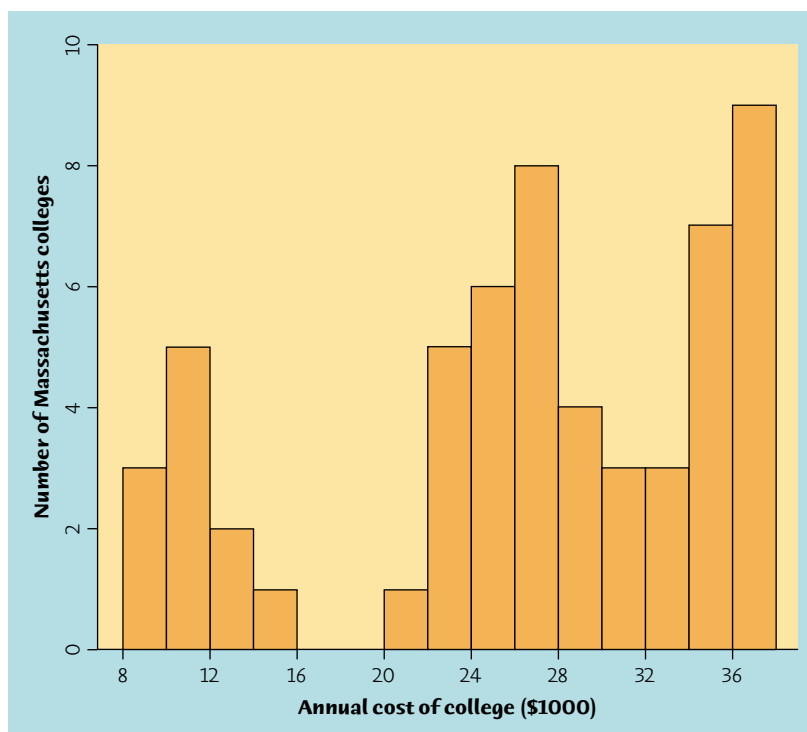


Figure 1.5 Histogram of the estimated costs (in thousands of dollars) for four-year colleges in Massachusetts. The two clusters distinguish public from private institutions.

The overall shape of a distribution is important information about a variable. Some types of data regularly produce distributions that are symmetric or skewed. For example, the sizes of living things of the same species (like lengths of crickets) tend to be symmetric. Data on incomes (whether of individuals, companies, or nations) are usually strongly skewed to the right. There are many moderate incomes, some large incomes, and a few very large incomes. Many distributions have irregular shapes that are neither symmetric nor skewed. Some data show other patterns, such as the clusters in Figure 1.5. Use your eyes and describe what you see.

APPLY YOUR KNOWLEDGE

- 1.6 Sports car fuel economy.** Table 1.2 (page 12) gives data on the fuel economy of 2002 model sports cars. Your histogram (Exercise 1.5) shows an extreme high outlier. This is the Honda Insight, a hybrid gas-electric car that is quite different from the others listed. Make a new histogram of highway mileage, leaving out the Insight. Classes that are about 2 miles per gallon wide work well.
- Describe the main features (shape, center, spread, outliers) of the distribution of highway mileage.
 - The government imposes a “gas guzzler” tax on cars with low gas mileage. Which of these cars do you think may be subject to the gas guzzler tax?
- 1.7 College costs.** Describe the center (midpoint) and spread (smallest to largest) of the distribution of Massachusetts college costs in Figure 1.5. An overall description works poorly because of the clusters. A better description gives the center and spread of each cluster (public and private colleges) separately. Do this.

Quantitative variables: stemplots

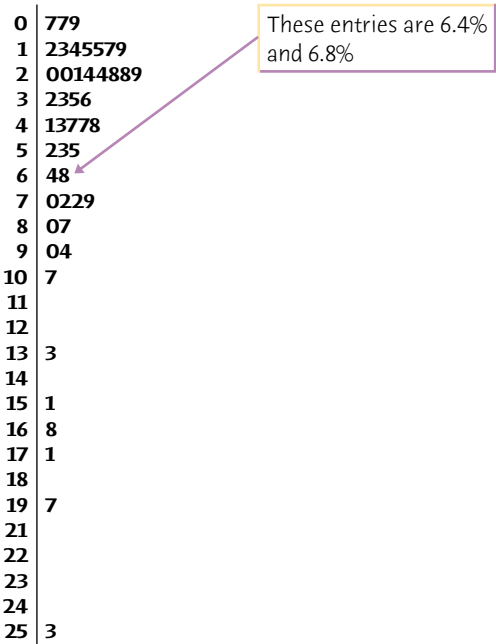
Histograms are not the only graphical display of distributions. For small data sets, a *stemplot* is quicker to make and presents more detailed information.

STEMPLOT

To make a **stemplot**:

- Separate each observation into a **stem**, consisting of all but the final (rightmost) digit, and a **leaf**, the final digit. Stems may have as many digits as needed, but each leaf contains only a single digit.
- Write the stems in a vertical column with the smallest at the top, and draw a vertical line at the right of this column.
- Write each leaf in the row to the right of its stem, in increasing order out from the stem.

Figure 1.6 Stemplot of the percents of Hispanic residents in the states. Each stem is a percent and leaves are tenths of a percent.



The vital few?

Skewed distributions can show us where to concentrate our efforts. Ten percent of the cars on the road account for half of all carbon dioxide emissions. A histogram of CO₂ emissions would show many cars with small or moderate values and a few with very high values. Cleaning up or replacing these cars would reduce pollution at a cost much lower than that of programs aimed at all cars. Statisticians who work at improving quality in industry make a principle of this: distinguish “the vital few” from “the trivial many.”

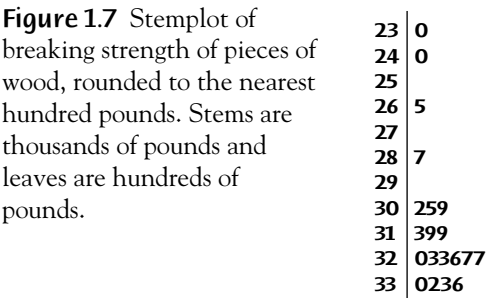
EXAMPLE 1.7 Making a stemplot

For the percents of Hispanic residents in Table 1.1, take the whole-number part of the percent as the stem and the final digit (tenths) as the leaf. The Massachusetts entry, 6.8%, has stem 6 and leaf 8. Wyoming, at 6.4%, places leaf 4 on the same stem. These are the only observations on this stem. We then arrange the leaves in order, as 48, so that 6 | 48 is one row in the stemplot. Figure 1.6 is the complete stemplot for the data in Table 1.1. To save space, we left out California, Texas, and New Mexico, which have stems 32 and 42.

A stemplot looks like a histogram turned on end. Compare the stemplot in Figure 1.6 with the histogram of the same data in Figure 1.3. Both show a single-peaked distribution that is strongly right-skewed and has some observations that we would probably call high outliers (three of these are left out of Figure 1.6). You can choose the classes in a histogram. The classes (the stems) of a stemplot are given to you. Figure 1.6 has more stems than there are classes in Figure 1.3. So histograms are more flexible. But the stemplot, unlike the histogram, preserves the actual value of each observation. Stemplots work well for small sets of data. Use a histogram to display larger data sets, like the 947 Iowa Test scores in Figure 1.4.

EXAMPLE 1.8 Pulling wood apart

Student engineers learn that although handbooks give the strength of a material as a single number, in fact the strength varies from piece to piece. A vital lesson in all fields of study is that “variation is everywhere.” Here are data from a typical student



laboratory exercise: the load in pounds needed to pull apart pieces of Douglas fir 4 inches long and 1.5 inches square.

33,190	31,860	32,590	26,520	33,280
32,320	33,020	32,030	30,460	32,700
23,040	30,930	32,720	33,650	32,340
24,050	30,170	31,300	28,730	31,920

We want to make a stemplot to display the distribution of breaking strength. To avoid many stems with only one leaf each, first **round** the data to the nearest hundred pounds. The rounded data are

rounding

332	319	326	265	333	323	330	320	305	327
230	309	327	336	323	240	302	313	287	319

Now it is easy to make a stemplot with the first two digits (thousands of pounds) as stems and the third digit (hundreds of pounds) as leaves. Figure 1.7 is the stemplot. The distribution is skewed to the left, with midpoint around 320 (32,000 pounds) and spread from 230 to 336.

You can also **split stems** to double the number of stems when all the leaves would otherwise fall on just a few stems. Each stem then appears twice. Leaves 0 to 4 go on the upper stem, and leaves 5 to 9 go on the lower stem. If you split the stems in the stemplot of Figure 1.7, for example, the 32 and 33 stems become

splitting stems

32	033
32	677
33	023
33	6

Rounding and splitting stems are matters for judgment, like choosing the classes in a histogram. The wood strength data require rounding but don't require splitting stems.

APPLY YOUR KNOWLEDGE

1.8 Students' attitudes. The Survey of Study Habits and Attitudes (SSHA) is a psychological test that evaluates college students' motivation, study habits, and attitudes toward school. A private college gives the SSHA

to 18 of its incoming first-year women students. Their scores are

154 109 137 115 152 140 154 178 101
103 126 126 137 165 165 129 200 148

Make a stemplot of these data. The overall shape of the distribution is irregular, as often happens when only a few observations are available. Are there any outliers? About where is the center of the distribution (the score with half the scores above it and half below)? What is the spread of the scores (ignoring any outliers)?

1.9 Alternative stemplots. Return to the Hispanics data in Table 1.1 and Figure 1.6. Round each state’s percent Hispanic to the nearest whole percent. Make a stemplot using tens of percents as stems and percents as leaves. All of the leaves fall on just five stems, 0, 1, 2, 3, and 4. Make another stemplot using split stems to increase the number of classes. With Figure 1.6, you now have three stemplots of the Hispanics data. Which do you prefer? Why?

Time plots

Many variables are measured at intervals over time. We might, for example, measure the height of a growing child or the price of a stock at the end of each month. In these examples, our main interest is change over time. To display change over time, make a *time plot*.

TIME PLOT

A **time plot** of a variable plots each observation against the time at which it was measured. Always put time on the horizontal scale of your plot and the variable you are measuring on the vertical scale. Connecting the data points by lines helps emphasize any change over time.

EXAMPLE 1.9 More on the cost of college

How have college tuition and fees changed over time? Table 1.3 gives the average tuition and fees paid by college students at four-year colleges, both public and private, from the 1971–1972 academic year to the 2001–2002 academic year. To compare dollar amounts across time, we must adjust for the changing buying power of the dollar. Table 1.3 gives tuition in *real dollars*, dollars that have constant buying power.⁷ Average tuition in real dollars goes up only when the actual tuition rises by more than the overall cost of living. Figure 1.8 is a time plot of both public and private tuition.

TABLE 1.3 Average college tuition and fees, 1971–1972 to 2001–2002, in real dollars

Year	Private colleges	Public colleges	Year	Private colleges	Public colleges	Year	Private colleges	Public colleges
1971	7,851	1,622	1982	8,389	1,865	1992	13,012	2,907
1972	7,870	1,688	1983	8,882	2,002	1993	13,362	3,077
1973	7,572	1,667	1984	9,324	2,061	1994	13,830	3,192
1974	7,255	1,481	1985	9,984	2,150	1995	14,035	3,229
1975	7,272	1,386	1986	10,502	2,051	1996	14,514	3,323
1976	7,664	1,866	1987	10,799	2,275	1997	15,128	3,414
1977	7,652	1,856	1988	11,723	2,311	1998	15,881	3,506
1978	7,665	1,783	1989	12,110	2,371	1999	16,289	3,529
1979	7,374	1,687	1990	12,380	2,529	2000	16,456	3,535
1980	7,411	1,647	1991	12,601	2,706	2001	17,123	3,754
1981	7,758	1,714						

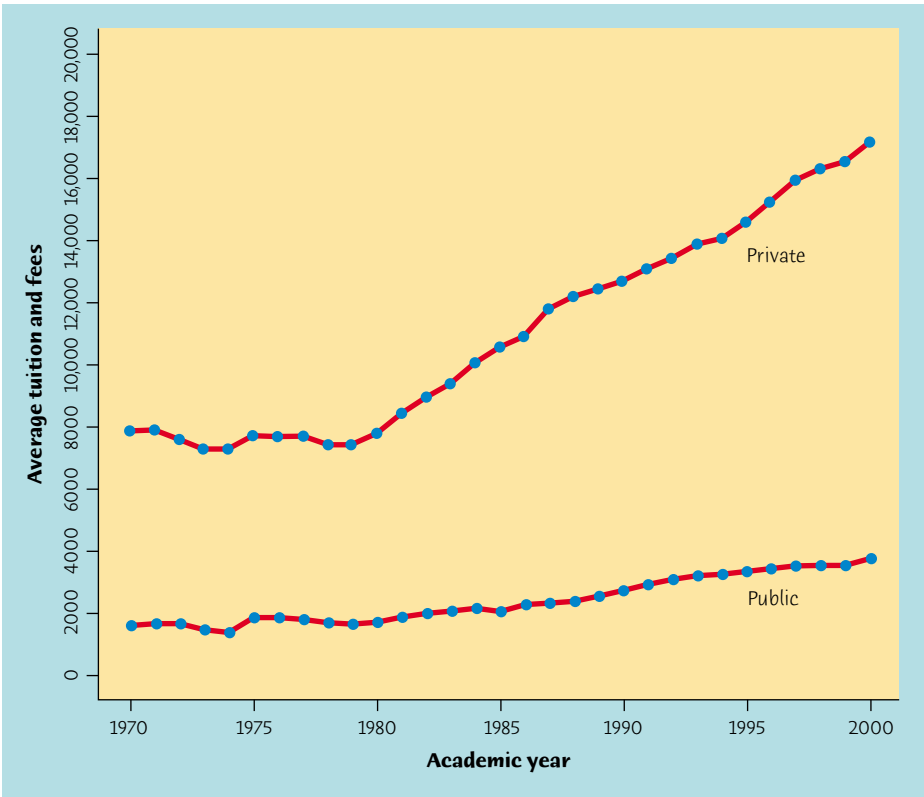


Figure 1.8 Time plot of the average tuition paid by students at public and private colleges for academic years 1970–1971 to 2001–2002.

trend

When you examine a time plot, look once again for an overall pattern and for strong deviations from the pattern. One common overall pattern is a **trend**, a long-term upward or downward movement over time. Figure 1.8 shows an upward trend in real college tuition costs, with no striking deviations such as short-term drops. It also shows that, beginning around 1980, private colleges raised tuition faster than public institutions, increasing the gap in costs between the two types of colleges.

time series
cross-sectional

Figures 1.5 and 1.8 both give information about college costs. The data for the time plot in Figure 1.8 are **time series data** that show the change in average tuition over time. The data for the histogram in Figure 1.5 are **cross-sectional data** that show the variation in costs (in one state) at a fixed time.

APPLY YOUR KNOWLEDGE



(Lester Lefkowitz/Corbis)

1.10 Vanishing landfills. The bar graphs in Figure 1.2 give cross-sectional data on municipal solid waste in 2000. Garbage that is not recycled is buried in landfills. Here are time series data that emphasize the need for recycling: the number of landfills operating in the United States in the years 1988 to 2000.⁸

Year	Landfills	Year	Landfills	Year	Landfills
1988	7924	1993	4482	1997	2514
1989	7379	1994	3558	1998	2314
1990	6326	1995	3197	1999	2216
1991	5812	1996	3091	2000	1967
1992	5386				

Make a time plot of these data. Describe the trend that your plot shows. Why does the trend emphasize the need for recycling?

Chapter 1 SUMMARY

A data set contains information on a number of **individuals**. Individuals may be people, animals, or things. For each individual, the data give values for one or more **variables**. A variable describes some characteristic of an individual, such as a person’s height, gender, or salary.

Some variables are **categorical** and others are **quantitative**. A categorical variable places each individual into a category, like male or female. A quantitative variable has numerical values that measure some characteristic of each individual, like height in centimeters or salary in dollars per year.

Exploratory data analysis uses graphs and numerical summaries to describe the variables in a data set and the relations among them.

The **distribution** of a variable describes what values the variable takes and how often it takes these values.

To describe a distribution, begin with a graph. **Bar graphs** and **pie charts** describe the distribution of a categorical variable. **Histograms** and **stemplots** graph the distribution of a quantitative variable.

When examining any graph, look for an **overall pattern** and for notable **deviations** from the pattern.

Shape, center, and spread describe the overall pattern of a distribution. Some distributions have simple shapes, such as **symmetric** or **skewed**. Not all distributions have a simple overall shape, especially when there are few observations.

Outliers are observations that lie outside the overall pattern of a distribution. Always look for outliers and try to explain them.

When observations on a variable are taken over time, make a **time plot** that graphs time horizontally and the values of the variable vertically. A time plot can reveal **trends** or other changes over time.

Chapter 1 EXERCISES

1.11 Car colors in Japan. Exercise 1.3 (page 7) gives data on the most popular colors for motor vehicles made in North America. Here are similar data for 2001 model year vehicles made in Japan:⁹

Color	Percent
Gray	43%
White	35%
Black	8%
Blue	7%
Red	4%
Green	2%

What percent of Japanese vehicles have other colors? Make a graph of these data. What are the most important differences between choice of vehicle color in Japan and North America?

1.12 Deaths among young people. The number of deaths among persons aged 15 to 24 years in the United States in 2000 due to the leading causes of death for this age group were: accidents, 13,616; homicide, 4796; suicide, 3877; cancer, 1668; heart disease, 931; congenital defects, 425.¹⁰

- (a) Make a bar graph to display these data.
(b) What additional information do you need to make a pie chart?

1.13 Athletes’ salaries. Here is a small part of a data set that describes Major League Baseball players as of opening day of the 2002 season:

Player	Team	Position	Age	Salary
⋮				
Sosa, Jorge	Devil Rays	Pitcher	24	200,000
Sosa, Sammy	Cubs	Outfield	33	15,000,000
Speier, Justin	Rockies	Pitcher	28	310,000
Spivey, Junior	Diamondbacks	Infield	27	215,000
⋮				

- (a) What individuals does this data set describe?
- (b) In addition to the player’s name, how many variables does the data set contain? Which of these variables are categorical and which are quantitative?
- (c) Based on the data in the table, what do you think are the units of measurement for each of the quantitative variables?

1.14 Mutual funds. Here is information on several Vanguard Group mutual funds:

Fund	Number of stocks held	Largest holding	Annual return (10 years)
500 Index Fund	508	General Electric	10.01%
Equity Income Fund	167	ExxonMobil	11.96%
Health Care Fund	128	Pharmacia	20.27%
International Value Fund	84	Mazda Motor	5.04%
Precious Metals Fund	26	Barrick Gold	2.50%

In addition to the fund name, how many variables are recorded for each fund? Which variables are categorical and which are quantitative?

1.15 Reading a pie chart. Figure 1.9 is a pie chart prepared by the Census Bureau to show the origin of the 35.3 million Hispanics in the United States, according to the 2000 census.¹¹ About what percent of Hispanics are Mexican? Puerto Rican? You see that it is hard to read numbers from a pie chart. Bar graphs are much easier to use.

1.16 Do adolescent girls eat fruit? We all know that fruit is good for us. Many of us don’t eat enough. Figure 1.10 is a histogram of the number of servings of fruit per day claimed by 74 seventeen-year-old girls in a study in Pennsylvania.¹² Describe the shape, center, and spread of this distribution. What percent of these girls ate fewer than two servings per day?

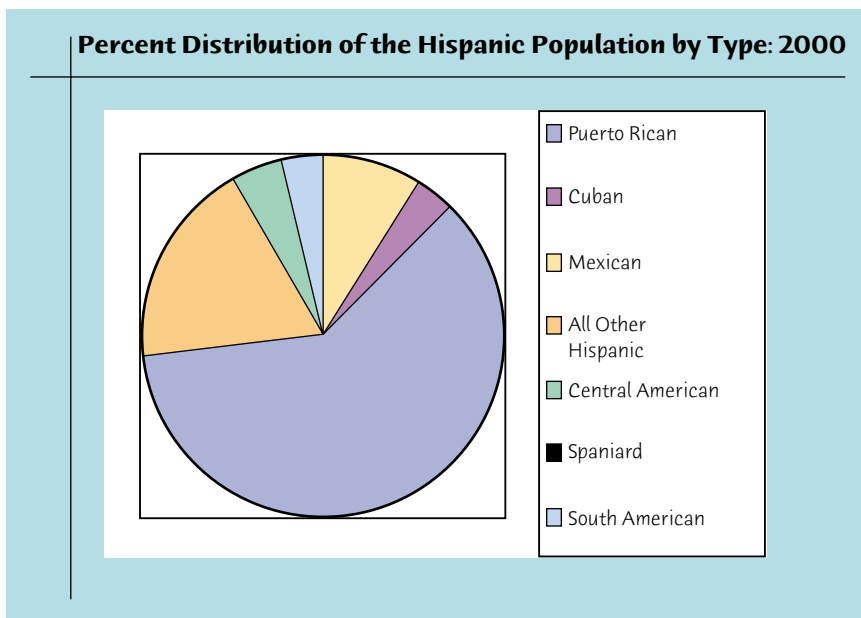


Figure 1.9 Pie chart of the origins of Hispanic residents of the United States, for Exercise 1.15. (Data from U.S. Census Bureau.)

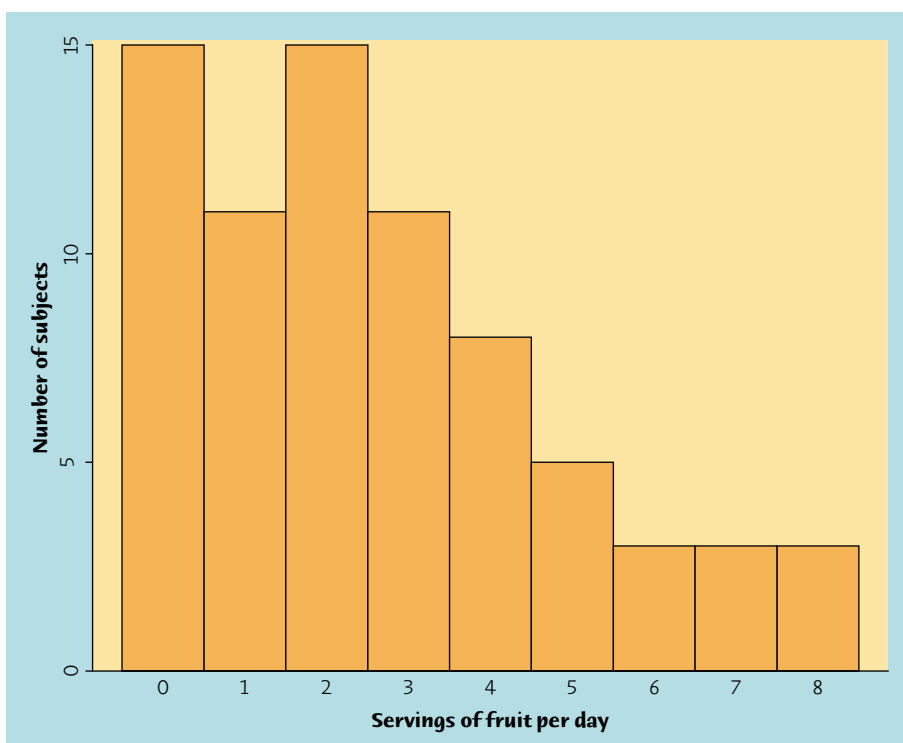


Figure 1.10 The distribution of fruit consumption in a sample of 74 seventeen-year-old girls, for Exercise 1.16.

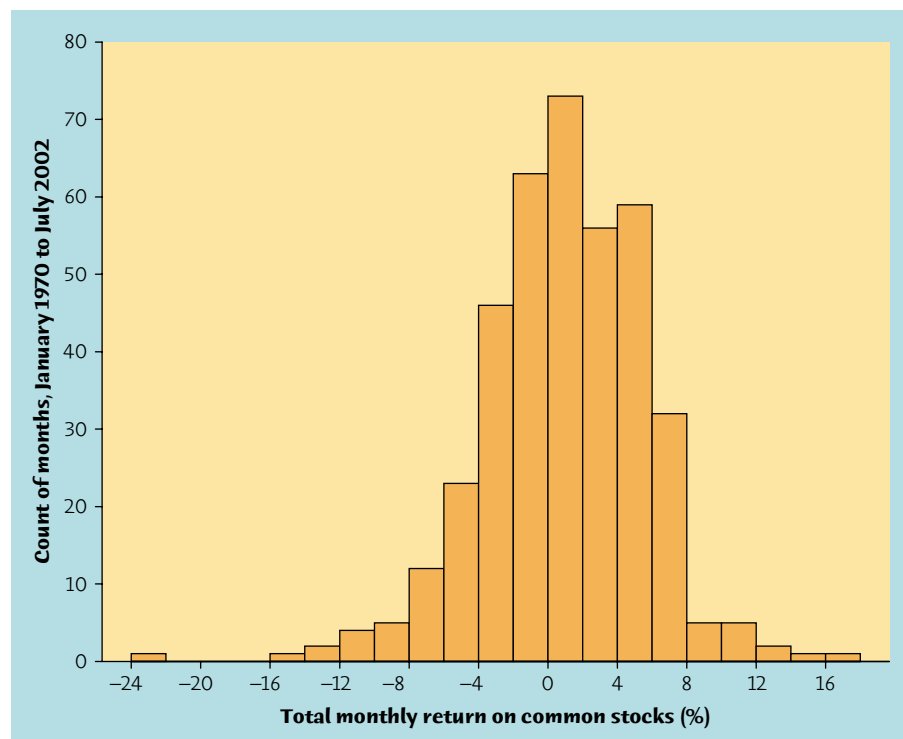


Figure 1.11 The distribution of monthly percent returns on U.S. common stocks from January 1970 to July 2002, for Exercise 1.17.

1.17 Returns on common stocks. The return on a stock is the change in its market price plus any dividend payments made. Total return is usually expressed as a percent of the beginning price. Figure 1.11 is a histogram of the distribution of the monthly returns for all stocks listed on U.S. markets from January 1970 to July 2002 (391 months).¹³ The low outlier is the market crash of October 1987, when stocks lost more than 22% of their value in one month.

- Describe the overall shape of the distribution of monthly returns.
- What is the approximate center of this distribution? (For now, take the center to be the value with roughly half the months having lower returns and half having higher returns.)
- Approximately what were the smallest and largest monthly returns, leaving out the outlier? (This describes the spread of the distribution.)
- A return less than zero means that stocks lost value in that month. About what percent of all months had returns less than zero?

1.18 Weight of newborns. Here is the distribution of the weight at birth for all babies born in the United States in 1999:¹⁴

Weight	Count	Weight	Count
Less than 500 grams	5,912	3,000 to 3,499 grams	1,470,019
500 to 999 grams	22,815	3,500 to 3,999 grams	1,137,401
1,000 to 1,499 grams	28,750	4,000 to 4,499 grams	332,863
1,500 to 1,999 grams	59,531	4,500 to 4,999 grams	53,751
2,000 to 2,499 grams	184,175	5,000 to 5,499 grams	6,069
2,500 to 2,999 grams	653,327		

- (a) For comparison with other years and with other countries, we prefer a histogram of the *percents* in each weight class rather than the counts. Explain why.
- (b) Make a histogram of the distribution, using percents on the vertical scale.
- (c) A “low-birth-weight” baby is one weighing less than 2,500 grams. Low birth weight is tied to many health problems. What percent of all births were low birth weight babies?

1.19 Marijuana and traffic accidents. Researchers in New Zealand interviewed 907 drivers at age 21. They had data on traffic accidents and they asked their subjects about marijuana use. Here are data on the numbers of accidents caused by these drivers at age 19, broken down by marijuana use at the same age:¹⁵

	Marijuana use per year			
	Never	1–10 times	11–50 times	51+ times
Drivers	452	229	70	156
Accidents caused	59	36	15	50

- (a) Explain carefully why a useful graph must compare *rates* (accidents per driver) rather than counts of accidents in the four marijuana use classes.
- (b) Make a graph that displays the accident rate for each class. What do you conclude? (You can’t conclude that marijuana use *causes* accidents, because risk-takers are more likely both to drive aggressively and to use marijuana.)

1.20 Lions feeding. Feeding at a carcass leads to competition among lions. Ecologists collected data on feeding contests in Serengeti National Park, Tanzania.¹⁶ In each contest, a lion feeding at a carcass is challenged by another lion seeking to take its place. Who wins these contests tells us something about lion society. The following table presents data on

396 contests between an adult lion (female or male) and an opponent of a different class:

Opponent	Adult male		Adult female	
	Contests	Male wins	Contests	Female wins
Adult female	136	113		
Subadult	8	8	53	20
Yearling	13	12	8	5
Cub	25	22	153	115

Make separate graphs for males and females that compare their success against different classes of opponent. Explain why you decided to make the type of graph you chose. Then describe the most important differences between the behavior of female and male lions in feeding contests.

- 1.21 **Name that variable.** Figure 1.12 displays four histograms without axis markings. They are the distributions of these four variables:¹⁷
1. The gender of the students in a large college course, recorded as 0 for male and 1 for female.
 2. The heights of the students in the same class.
 3. The handedness of students in the class, recorded as 0 for right-handed and 1 for left-handed.
 4. The lengths of words used in Shakespeare’s plays.

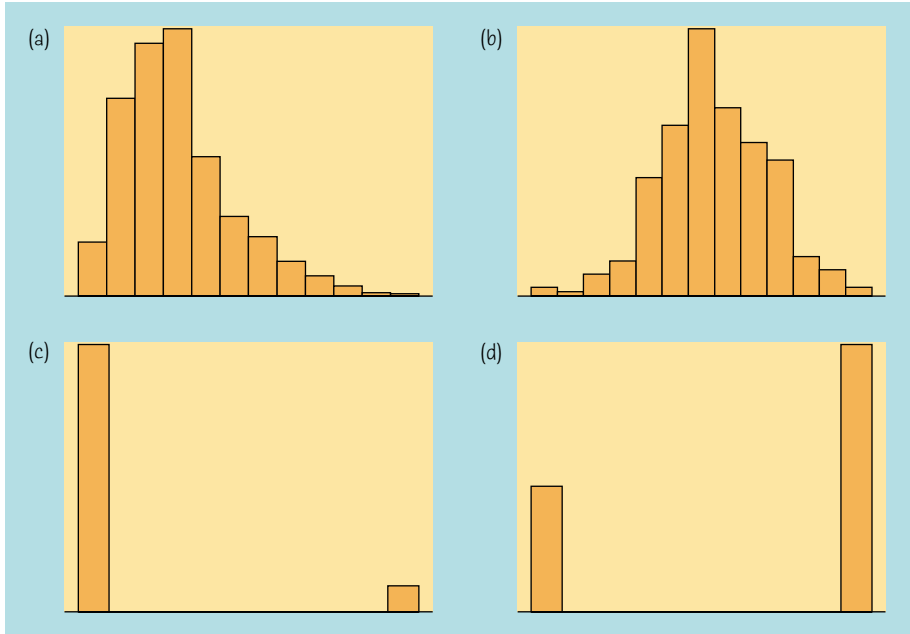


Figure 1.12 Histograms of four distributions, for Exercise 1.21.

TABLE 1.4 Percent of state residents living in poverty, 2000

State	Percent	State	Percent	State	Percent
Alabama	14.6	Maryland	7.3	Pennsylvania	9.9
Connecticut	7.6	Massachusetts	10.2	Rhode Island	10.0
Delaware	9.8	Michigan	10.2	South Carolina	11.9
Florida	12.1	Mississippi	15.5	Tennessee	13.3
Georgia	12.6	New Hampshire	7.4	Vermont	10.1
Illinois	10.5	New Jersey	8.1	Virginia	8.1
Indiana	8.2	New York	14.7	West Virginia	15.8
Kentucky	12.5	North Carolina	13.2	Wisconsin	8.8
Maine	9.8	Ohio	11.1		

Identify which distribution each histogram describes. Explain the reasoning behind your choices.

- 1.22 Dates on coins.** Sketch a histogram for a distribution that is skewed to the left. Suppose that you and your friends emptied your pockets of coins and recorded the year marked on each coin. The distribution of dates would be skewed to the left. Explain why.
- 1.23 Poverty in the states.** Table 1.4 gives the percents of people living below the poverty line in the 26 states east of the Mississippi River.¹⁸ Make a stemplot of these data. Is the distribution roughly symmetric, skewed to the right, or skewed to the left? Which states (if any) are outliers?
- 1.24 Split the stems.** Make another stemplot of the poverty data in Table 1.4, splitting the stems to double the number of classes. Do you prefer this stemplot or that from the previous exercise? Why?
- 1.25 Babe Ruth’s home runs.** Here are the numbers of home runs that Babe Ruth hit in his 15 years with the New York Yankees, 1920 to 1934:

54 59 35 41 46 25 47 60 54 46 49 46 41 34 22

Make a stemplot for these data. Is the distribution roughly symmetric, clearly skewed, or neither? About how many home runs did Ruth hit in a typical year? Is his famous 60 home runs in 1927 an outlier?

- 1.26 Back-to-back stemplot.** A leading contemporary home run hitter is Mark McGwire, who retired after the 2001 season. Here are McGwire’s home run counts for 1987 to 2001:

49 32 33 39 22 42 9 9 39 52 58 70 65 32 29

A *back-to-back stemplot* helps us compare two distributions. Write the stems as usual, but with a vertical line both to their left and to their right. On the right, put leaves for Babe Ruth (see the previous exercise). On the left, put leaves for McGwire. Arrange the leaves on each stem in



(National Baseball Hall of Fame and Museum, Inc., Cooperstown, N.Y.)

TABLE 1.5 Women’s winning times (minutes) in the Boston Marathon

Year	Time	Year	Time	Year	Time
1972	190	1982	150	1992	144
1973	186	1983	143	1993	145
1974	167	1984	149	1994	142
1975	162	1985	154	1995	145
1976	167	1986	145	1996	147
1977	168	1987	146	1997	146
1978	165	1988	145	1998	143
1979	155	1989	144	1999	143
1980	154	1990	145	2000	146
1981	147	1991	144	2001	144
				2002	141

increasing order out from the stem. Now write a brief comparison of Ruth and McGwire as home run hitters. McGwire was injured in 1993 and there was a baseball strike in 1994. How do these events appear in the data?

1.27 The Boston Marathon. Women were allowed to enter the Boston Marathon in 1972. The times (in minutes, rounded to the nearest minute) for the winning woman from 1972 to 2002 appear in Table 1.5. In 2002, Margaret Okayo of Kenya set a new women’s record for the race of 2 hours, 20 minutes, and 43 seconds.

- (a) Make a time plot of the winning times.
- (b) Give a brief description of the pattern of Boston Marathon winning times over these years. Has the rate of improvement in times slowed in recent years?

1.28 Watch those scales! The impression that a time plot gives depends on the scales you use on the two axes. If you stretch the vertical axis and compress the time axis, change appears to be more rapid. Compressing the vertical axis and stretching the time axis make change appear slower. Make two more time plots of the data on public college tuition in Table 1.3, one that makes tuition appear to increase very rapidly and one that shows only a gentle increase. The moral of this exercise is: pay close attention to the scales when you look at a time plot.

1.29 Where are the doctors? Table 1.6 gives the number of medical doctors per 100,000 people in each state.¹⁹

- (a) Why is the number of doctors per 100,000 people a better measure of the availability of health care than a simple count of the number of doctors in a state?
- (b) Make a graph that displays the distribution of doctors per 100,000 people. Write a brief description of the distribution. Are there any outliers? If so, can you explain them?

TABLE 1.6 Medical doctors per 100,000 people, by state (1999)

State	Doctors	State	Doctors	State	Doctors
Alabama	200	Louisiana	251	Ohio	237
Alaska	170	Maine	232	Oklahoma	167
Arizona	203	Maryland	379	Oregon	227
Arkansas	192	Massachusetts	422	Pennsylvania	293
California	248	Michigan	226	Rhode Island	339
Colorado	244	Minnesota	254	South Carolina	213
Connecticut	361	Mississippi	164	South Dakota	188
Delaware	238	Missouri	232	Tennessee	248
Florida	243	Montana	191	Texas	205
Georgia	211	Nebraska	221	Utah	202
Hawaii	269	Nevada	177	Vermont	313
Idaho	155	New Hampshire	234	Virginia	243
Illinois	263	New Jersey	301	Washington	237
Indiana	198	New Mexico	214	West Virginia	219
Iowa	175	New York	395	Wisconsin	232
Kansas	204	North Carolina	237	Wyoming	172
Kentucky	212	North Dakota	224	D.C.	758

1.30 Orange prices. Figure 1.13 is a time plot of the average price of fresh oranges each month during the decade from 1992 to 2002.²⁰ The prices are “index numbers” given as percents of the average price during 1982 to 1984.

- (a) The most notable pattern in this time plot is **seasonal variation**, regular up-and-down movements that occur at about the same time each year. Why should we expect the price of fresh oranges to show seasonal variation?
- (b) Is there a longer-term trend visible under the seasonal variation? If so, describe it.

seasonal variation

Chapter 1 MEDIA EXERCISES

The Internet is now the first place to look for data. Many of the data sets in this chapter were found online. Exercises 1.31 to 1.33 illustrate some places to find data on the Web.

1.31 No place like home? The Census Bureau Web site, www.census.gov, is the mother lode of data about America and Americans. On the home page, you can select a state and then a county within that state. Select your home county and then the county in which your school is located. For each county, record the population, the percent population change in the decade 1990 to 2000, and the percents of Asian, black, and white residents. Also calculate the number of people age 25 and over with



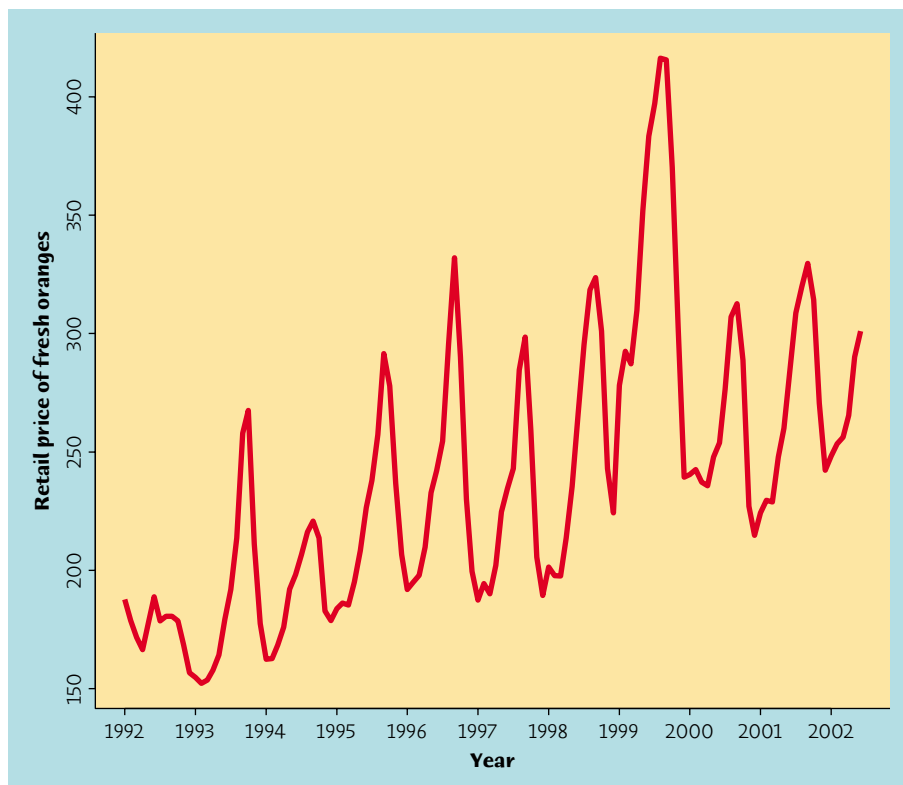


Figure 1.13 Time plot of the monthly average price of fresh oranges, for Exercise 1.30.

college degrees as a percent of the county's population. What are the main differences between your home county and your school's county?



1.32 Canada and the U.S.A. What are the very latest estimates of the populations of Canada and the United States? Find them at Statistics Canada, www.statcan.ca, and the U.S. Census Bureau, www.census.gov. What are the latest unemployment rates in the two countries? Canada, like most nations, has a single statistical agency, so the Statistics Canada site is the place to look. In the United States, statistical agencies are attached to many government departments. For unemployment data, go to the Bureau of Labor Statistics, www.bls.gov.



1.33 Current issues. The University of Chicago's National Opinion Research Center carries out polls of public opinion on many issues, as well as other studies. At www.norc.org you will find recent reports and press releases. For example, late in 2002 the site featured reports on opinions about owning and regulating guns, reactions to the September 11, 2001, attack on the World Trade Center, and rating hospitals. Choose a topic that interests you among those on the home page and browse until you find data that you think shed light on the

topic. (For example, 76.9% of those polled in 2001 wanted mandatory registration of handguns.) Report the data and say why you think they are important.

1.34 How histograms behave. The data set menu that accompanies the *One-Variable Statistical Calculator* applet includes the data on Hispanics in the states from Table 1.1. Choose these data, then click on the “Histogram” tab to see a histogram.



- How many classes does the applet choose to use? (You can click on the graph outside the bars to get a count of classes.)
- Click on the graph and drag to the left. What is the smallest number of classes you can get? What are the lower and upper bounds of each class? (Click on the bar to find out.) Make a rough sketch of this histogram.
- Click and drag to the right. What is the greatest number of classes you can get? How many observations does the largest class have?
- You see that the choice of classes changes the appearance of a histogram. Drag back and forth until you get the histogram you think best displays the distribution. How many classes did you use?

1.35 Choices in a stemplot. The data set menu that accompanies the *One-Variable Statistical Calculator* applet includes the data on Hispanics in the states from Table 1.1. Choose these data, then click on the “Stemplot” tab to see a stemplot.



- The stemplot looks quite different from that in Figure 1.6. Make a copy of this stemplot, and explain carefully the reason for the difference.
- Figure 1.6 has 26 stems and would have 43 stems if we extended it to include New Mexico. The applet’s plot has many fewer. Check the “Split stems” box to increase the number of stems used by the applet. Make a copy of this stemplot as well. You now have three stemplots for these data. Which do you prefer, and why?

1.36 Acorns. How big are acorns? It depends on the species of oak tree that produces them. The EESEE story “Acorn Size and Oak Tree Range” contains data on the average size (in cubic centimeters) of acorns from 39 species of oaks. Make a stemplot of the acorn size data. Describe the distribution carefully (shape, center, spread, outliers).



1.37 Eruptions of Old Faithful. The EESEE story “Is Old Faithful Faithful?” contains data on eruptions of the famous Old Faithful geyser in Yellowstone National Park. The variable named “Duration” records how long 299 of these eruptions lasted, in minutes. Use your software to make a histogram of the durations. The shape of the distribution is distinctive and interesting. Describe the shape, center, and spread of the distribution.



CHAPTER 2



(Tony Arruza/Corbis)

In this chapter we cover...

Measuring center: the mean

Measuring center: the median

Comparing the mean and the median

Measuring spread: the quartiles

The five-number summary and boxplots

Measuring spread: the standard deviation

Choosing measures of center and spread

Using technology

Describing Distributions with Numbers

How much do people with a bachelor's degree (but no higher degree) earn? Here are the incomes of 15 such people, chosen at random by the Census Bureau in March 2002 and asked how much they earned in 2001.¹ Most people reported their incomes to the nearest thousand dollars, so we have rounded their responses to thousands of dollars.

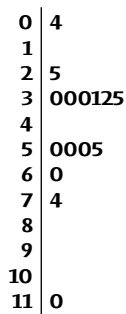
110 25 50 50 55 30 35 30 4 32 50 30 32 74 60

Figure 2.1 is a stemplot of these amounts. The distribution is irregular in shape, as is common when we have only a few observations. There is one high outlier, a person who made \$110,000. Our goal in this chapter is to describe with numbers the center and spread of this and other distributions.

Measuring center: the mean

The most common measure of center is the ordinary arithmetic average, or *mean*.

Figure 2.1 Stemplot of the earnings (in thousands of dollars) of 15 people chosen at random from all people with a bachelor's degree but no higher degree.



THE MEAN $\bar{x}$

To find the **mean** of a set of observations, add their values and divide by the number of observations. If the n observations are $x_1, x_2, \dots, x_n$, their mean is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

or in more compact notation,

$$\bar{x} = \frac{1}{n} \sum x_i$$

The Σ (capital Greek sigma) in the formula for the mean is short for “add them all up.” The subscripts on the observations x_i are just a way of keeping the n observations distinct. They do not necessarily indicate order or any other special facts about the data. The bar over the x indicates the mean of all the x -values. Pronounce the mean $\bar{x}$ as “x-bar.” This notation is very common. When writers who are discussing data use $\bar{x}$ or $\bar{y}$, they are talking about a mean.

EXAMPLE 2.1 Earnings of college graduates

The mean earnings for our 15 college graduates are

$$\begin{aligned} \bar{x} &= \frac{x_1 + x_2 + \cdots + x_n}{n} \\ &= \frac{110 + 25 + \cdots + 60}{15} \\ &= \frac{666}{15} = 44.4, \text{ or } \$44,400 \end{aligned}$$

In practice, you can key the data into your calculator and hit the mean key. You don't have to actually add and divide. But you should know that this is what the calculator is doing.

If we leave out the one high income, \$110,000, the mean for the remaining 14 people is \$39,700. The lone outlier raises the mean income of the group by \$4700.

resistant measure

Example 2.1 illustrates an important fact about the mean as a measure of center: it is sensitive to the influence of a few extreme observations. These may be outliers, but a skewed distribution that has no outliers will also pull the mean toward its long tail. Because the mean cannot resist the influence of extreme observations, we say that it is not a **resistant measure** of center.

APPLY YOUR KNOWLEDGE

- 2.1 Sports car gas mileage.** Table 1.2 (page 12) gives the gas mileages for the 22 two-seater cars listed in the government’s fuel economy guide.
- (a) Find the mean highway gas mileage from the formula for the mean. Then enter the data into your calculator and use the calculator’s $\bar{x}$ button to obtain the mean. Verify that you get the same result.
 - (b) The Honda Insight is an outlier that doesn’t belong with the other cars. Use your calculator to find the mean of the 21 cars that remain if we leave out the Insight. How does the outlier change the mean?

Measuring center: the median

In Chapter 1, we used the midpoint of a distribution as an informal measure of center. The *median* is the formal version of the midpoint, with a specific rule for calculation.

THE MEDIAN M

The **median** M is the midpoint of a distribution, the number such that half the observations are smaller and the other half are larger. To find the median of a distribution:

1. Arrange all observations in order of size, from smallest to largest.
2. If the number of observations n is odd, the median M is the center observation in the ordered list. Find the location of the median by counting $(n + 1)/2$ observations up from the bottom of the list.
3. If the number of observations n is even, the median M is the mean of the two center observations in the ordered list. The location of the median is again $(n + 1)/2$ from the bottom of the list.

Note that the formula $(n + 1)/2$ does *not* give the median, just the location of the median in the ordered list. Medians require little arithmetic, so they are easy to find by hand for small sets of data. Arranging even a moderate number of observations in order is very tedious, however, so that finding the median by hand for larger sets of data is unpleasant. Even simple calculators have an $\bar{x}$

button, but you will need to use software or a graphing calculator to automate finding the median.

EXAMPLE 2.2 Finding the median: odd n

What are the median earnings for our 15 college graduates? Here are the data arranged in order:

4 25 30 30 30 31 32 **35** 50 50 50 55 60 74 110

The count of observations $n = 15$ is odd. The bold 35 is the center observation in the ordered list, with 7 observations to its left and 7 to its right. This is the median, $M = 35$.

Because $n = 15$, our rule for the location of the median gives

$$\text{location of } M = \frac{n+1}{2} = \frac{16}{2} = 8$$

That is, the median is the 8th observation in the ordered list. It is faster to use this rule than to locate the center by eye.

EXAMPLE 2.3 Finding the median: even n

How much does the high outlier affect the median? Drop the 110 from the ordered list and find the median of the remaining 14 incomes. The data are

4 25 30 30 30 31 **32** **35** 50 50 50 55 60 74

There is no center observation, but there is a center pair. These are the bold 32 and 35 in the list, which have 6 observations to their left in the list and 6 to their right. The median is midway between these two observations:

$$M = \frac{32+35}{2} = 33.5$$

With $n = 14$, the rule for locating the median in the list gives

$$\text{location of } M = \frac{n+1}{2} = \frac{15}{2} = 7.5$$

The location 7.5 means “halfway between the 7th and 8th observations in the ordered list.” That agrees with what we found by eye.

Comparing the mean and the median

Examples 2.1 to 2.3 illustrate an important difference between the mean and the median. The single high income pulls the mean up by \$4700. It moves the median by only \$1500. The median, unlike the mean, is *resistant*. If the high earner’s income rose from 110 to 1100 (that is, from \$110,000 to \$1,100,000) the median would not change at all. The 1100 just counts as one observation above the center, no matter how far above the center it lies. The mean uses the actual value of each observation and so will chase a single large observation upward. The *Mean and Median* applet is an excellent way to compare the resistance of M and $\bar{x}$.



COMPARING THE MEAN AND THE MEDIAN

The mean and median of a symmetric distribution are close together. If the distribution is exactly symmetric, the mean and median are exactly the same. In a skewed distribution, the mean is farther out in the long tail than is the median.

Distributions of incomes are usually skewed to the right—there are many modest incomes and a few very high incomes. For example, the Census Bureau survey in March 2002 interviewed 16,018 people aged 25 to 65 who were in the labor force full-time in 2001 and who were college graduates but had only a bachelor’s degree. We used 15 of these 16,018 incomes to introduce the mean and median. The median income for the entire group was \$45,769. The mean of the same 16,018 incomes was much higher, \$59,852. Reports about incomes and other strongly skewed distributions usually give the median (“midpoint”) rather than the mean (“arithmetic average”). However, a county that is about to impose a tax of 1% on the incomes of its residents cares about the mean income, not the median. The tax revenue will be 1% of total income, and the total is the mean times the number of residents. The mean and median measure center in different ways, and both are useful.

APPLY YOUR KNOWLEDGE

- 2.2 Sports car gas mileage.** What is the median highway mileage for the 22 two-seater cars listed in Table 1.2 (page 12)? What is the median of the 21 cars that remain if we remove the Honda Insight? Compare the effect of the Insight on mean mileage (Exercise 2.1) and on the median mileage. What general fact about the mean and median does this comparison illustrate?
- 2.3 House prices.** The mean and median selling price of existing single-family homes sold in June 2002 were \$163,900 and \$210,900. Which of these numbers is the mean and which is the median? Explain how you know.
- 2.4 Barry Bonds.** The major league baseball single-season home run record is held by Barry Bonds of the San Francisco Giants, who hit 73 in 2001. Here are Bonds’s home run totals from 1986 (his first year) to 2002:
- 16 25 24 19 33 25 34 46 37 33 42 40 37 34 49 73 46
- Bonds’s record year is a high outlier. How do his career mean and median number of home runs change when we drop the record 73? What general fact about the mean and median does your result illustrate?

Measuring spread: the quartiles

The mean and median provide two different measures of the center of a distribution. But a measure of center alone can be misleading. The Census Bureau

reports that in 2001 the median income of American households was \$42,228. Half of all households had incomes below \$42,228, and half had higher incomes. The mean income of these same households was \$58,208. The mean is higher than the median because the distribution of incomes is skewed to the right. But the median and mean don't tell the whole story. The bottom 20% of households had incomes less than \$17,970 and households in the top 5% took in more than \$150,499.² We are interested in the *spread* or *variability* of incomes as well as their center. The simplest useful numerical description of a distribution consists of both a measure of center and a measure of spread.

One way to measure spread is to give the smallest and largest observations. For example, the incomes of our 15 college graduates range from \$4000 to \$110,000. These single observations show the full spread of the data, but they may be outliers. We can improve our description of spread by also looking at the spread of the middle half of the data. The *quartiles* mark out the middle half. Count up the ordered list of observations, starting from the smallest. The *first quartile* lies one-quarter of the way up the list. The *third quartile* lies three-quarters of the way up the list. In other words, the first quartile is larger than 25% of the observations, and the third quartile is larger than 75% of the observations. The second quartile is the median, which is larger than 50% of the observations. That is the idea of quartiles. We need a rule to make the idea exact. The rule for calculating the quartiles uses the rule for the median.

THE QUARTILES Q_1 AND Q_3

To calculate the **quartiles**:

1. Arrange the observations in increasing order and locate the median M in the ordered list of observations.
2. The **first quartile** Q_1 is the median of the observations whose position in the ordered list is to the left of the location of the overall median.
3. The **third quartile** Q_3 is the median of the observations whose position in the ordered list is to the right of the location of the overall median.

Here are examples that show how the rules for the quartiles work for both odd and even numbers of observations.

EXAMPLE 2.4 Finding the quartiles: odd n

Our sample of 15 incomes of college graduates, arranged in increasing order, is

4 25 30 30 30 31 32 35 50 50 50 55 60 74 110

There is an odd number of observations, so the median is the middle one, the bold 35 in the list. The first quartile is the median of the 7 observations to the left of the

median. This is the 4th of these 7 observations, so $Q_1 = 30$. If you want, you can use the recipe for the location of the median with $n = 7$:

$$\text{location of } Q_1 = \frac{n+1}{2} = \frac{7+1}{2} = 4$$

The third quartile is the median of the 7 observations to the right of the median, $Q_3 = 55$. The overall median is left out of the calculation of the quartiles when there is an odd number of observations.

Notice that the quartiles are resistant. For example, Q_3 would have the same value if the outlier were 1100 rather than 110.

EXAMPLE 2.5 Finding the quartiles: even n

Here, from the same government survey, are the earnings in 2001 of 16 randomly chosen people who have high school diplomas but no college. For convenience we have arranged the incomes in increasing order.

5 6 12 19 20 21 22 24 | 25 31 32 40 43 43 47 67

There is an even number of observations, so the median lies midway between the middle pair, the 8th and 9th in the list. Its value is $M = 24.5$. We have marked the location of the median by |. The first quartile is the median of the first 8 observations, because these are the observations to the left of the location of the median. Check that $Q_1 = 19.5$ and $Q_3 = 41.5$. When the number of observations is even, all the observations enter into the calculation of the quartiles.

Be careful when, as in these examples, several observations take the same numerical value. Write down all of the observations and apply the rules just as if they all had distinct values. Some software packages use a slightly different rule to find the quartiles, so computer results may be a bit different from your own work. Don't worry about this. The differences will always be too small to be important.

The five-number summary and boxplots

The smallest and largest observations tell us little about the distribution as a whole, but they give information about the tails of the distribution that is missing if we know only Q_1 , M , and Q_3 . To get a quick summary of both center and spread, combine all five numbers.

THE FIVE-NUMBER SUMMARY

The **five-number summary** of a distribution consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols, the five-number summary is

Minimum Q_1 M Q_3 Maximum

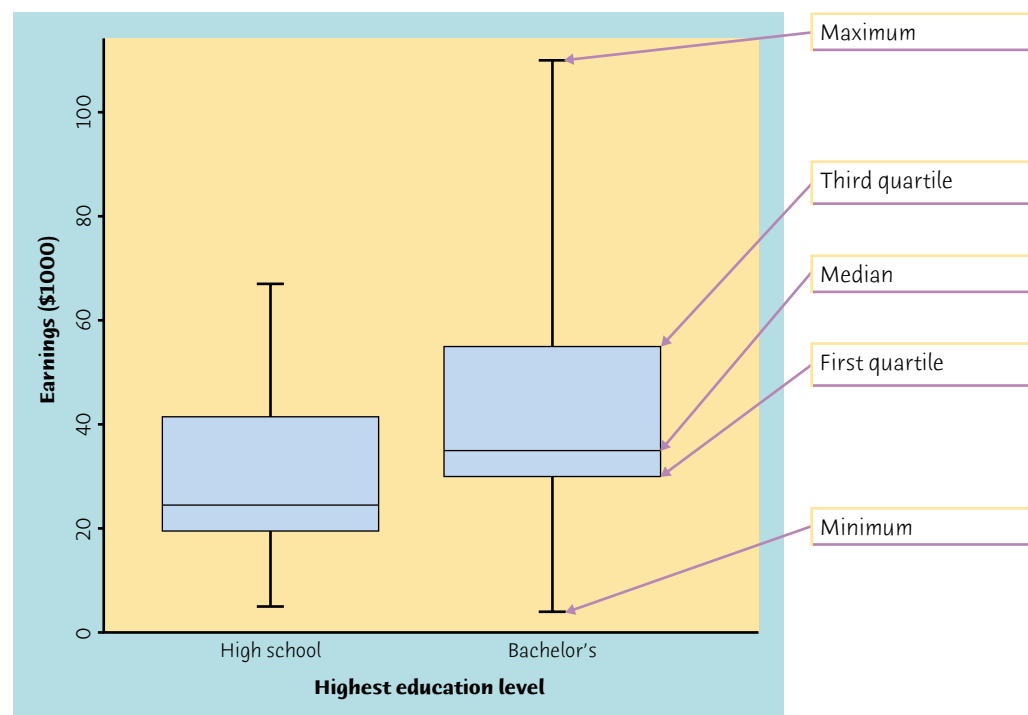


Figure 2.2 Side-by-side boxplots comparing the distributions of earnings for two levels of education.

These five numbers offer a reasonably complete description of center and spread. The five-number summaries from Examples 2.4 and 2.5 are

4 30 35 55 110

for people with a bachelor's degree, and

5 19.5 24.5 41.5 67

for people with only a high school diploma. The five-number summary of a distribution leads to a new graph, the *boxplot*. Figure 2.2 shows boxplots for the comparison of earnings.

BOXPLOT

A **boxplot** is a graph of the five-number summary.

- A central box spans the quartiles Q_1 and Q_3 .
- A line in the box marks the median M .
- Lines extend from the box out to the smallest and largest observations.

Because boxplots show less detail than histograms or stemplots, they are best used for side-by-side comparison of more than one distribution, as in Figure 2.2. Be sure to include a numerical scale in the graph. When you look at a boxplot, first locate the median, which marks the center of the distribution. Then look at the spread. The quartiles show the spread of the middle half of the data, and the extremes (the smallest and largest observations) show the spread of the entire data set. We see from Figure 2.2 that holders of a bachelor's degree as a group earn considerably more than people with no education beyond high school. For example, the first quartile for college graduates is higher than the median for high school grads. The spread of the middle half of incomes (the box in the boxplot) is roughly the same for both groups.

A boxplot also gives an indication of the symmetry or skewness of a distribution. In a symmetric distribution, the first and third quartiles are equally distant from the median. In most distributions that are skewed to the right, on the other hand, the third quartile will be farther above the median than the first quartile is below it. That is the case for both distributions in Figure 2.2. The extremes behave the same way, but remember that they are just single observations and may say little about the distribution as a whole.

APPLY YOUR KNOWLEDGE

2.5 Pulling wood apart. Example 1.8 (page 16) gives the breaking strengths of 20 pieces of Douglas fir.

- Give the five-number summary of the distribution of breaking strengths. (The stemplot, Figure 1.7, helps, because it arranges the data in order, but you should use the unrounded values in numerical work.)
- The stemplot shows that the distribution is skewed to the left. Does the five-number summary show the skew? Remember that only a graph gives a clear picture of the shape of a distribution.

2.6 Midsize car gas mileage. Table 2.1 gives the city and highway gas mileage for 36 midsize cars from the 2002 model year.³ There is one low outlier, the 12-cylinder Rolls-Royce. We wonder if midsize sedans get better mileage than sports cars.

- Find the five-number summaries for both city and highway mileage for the midsize cars in Table 2.1 and for the two-seater cars in Table 1.2. (Leave out the Honda Insight.)
- Make four side-by-side boxplots to display the summaries. Write a brief description of city versus highway mileage and two-seaters versus midsize cars.

2.7 How old are presidents? How old are presidents at their inauguration? Was Bill Clinton, at age 46, unusually young? Table 2.2 gives the data, the ages of all U.S. presidents when they took office.

- Make a stemplot of the distribution of ages. From the shape of the distribution, do you expect the median to be much less than the

TABLE 2.1 Gas mileage (miles per gallon) for 2002 model midsize cars

Model	City	Highway	Model	City	Highway
Acura 3.2TL	19	29	Mazda 626	22	28
Audi A6	19	25	Mercedes-Benz E320	20	28
Buick Century	20	29	Mercedes-Benz E430	17	24
Buick Regal	18	27	Mercury Sable	20	28
Cadillac Seville	18	27	Mitsubishi Diamante	18	25
Chevrolet Malibu	20	29	Mitsubishi Galant	21	28
Chevrolet Monte Carlo	21	32	Nissan Altima	23	29
Chrysler Sebring	21	30	Nissan Maxima	20	26
Daewoo Leganza	20	28	Oldsmobile Aurora	18	27
Honda Accord	23	30	Oldsmobile Intrigue	20	30
Hyundai Sonata	22	28	Pontiac Grand Prix	18	28
Infiniti I35	20	26	Rolls-Royce Silver Seraph	12	16
Infiniti Q45	17	25	Saab 9-3	12	29
Jaguar S-Type	18	25	Saab 9-5	20	30
Kia Optima	20	27	Saturn L100/200	24	33
Lexus ES300	21	29	Saturn L300	21	29
Lexus LS430	18	25	Toyota Camry	23	32
Lincoln LS	17	24	Volkswagen Passat	22	31

TABLE 2.2 Ages of the presidents at inauguration

President	Age	President	Age	President	Age
Washington	57	Lincoln	52	Hoover	54
J. Adams	61	A. Johnson	56	F. D. Roosevelt	51
Jefferson	57	Grant	46	Truman	60
Madison	57	Hayes	54	Eisenhower	61
Monroe	58	Garfield	49	Kennedy	43
J. Q. Adams	57	Arthur	51	L. B. Johnson	55
Jackson	61	Cleveland	47	Nixon	56
Van Buren	54	B. Harrison	55	Ford	61
W. H. Harrison	68	Cleveland	55	Carter	52
Tyler	51	McKinley	54	Reagan	69
Polk	49	T. Roosevelt	42	G. H. W. Bush	64
Taylor	64	Taft	51	Clinton	46
Fillmore	50	Wilson	56	G. W. Bush	54
Pierce	48	Harding	55		
Buchanan	65	Coolidge	51		

- mean, about the same as the mean, or much greater than the mean?
- (b) Find the mean and the five-number summary. Verify your expectation about the median.
- (c) What is the range of the middle half of the ages of new presidents? Was Bill Clinton in the youngest 25%?

Measuring spread: the standard deviation

The five-number summary is not the most common numerical description of a distribution. That distinction belongs to the combination of the mean to measure center and the *standard deviation* to measure spread. The standard deviation measures spread by looking at how far the observations are from their mean.

THE STANDARD DEVIATION s

The **variance** s^2 of a set of observations is an average of the squares of the deviations of the observations from their mean. In symbols, the variance of n observations $x_1, x_2, \dots, x_n$ is

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}$$

or, more compactly,

$$s^2 = \frac{1}{n - 1} \sum (x_i - \bar{x})^2$$

The **standard deviation** s is the square root of the variance s^2 :

$$s = \sqrt{\frac{1}{n - 1} \sum (x_i - \bar{x})^2}$$

In practice, use software or your calculator to obtain the standard deviation from keyed-in data. Doing an example step-by-step will help you understand how the variance and standard deviation work, however.

EXAMPLE 2.6 Calculating the standard deviation

A person's metabolic rate is the rate at which the body consumes energy. Metabolic rate is important in studies of weight gain, dieting, and exercise. Here are the metabolic rates of 7 men who took part in a study of dieting. (The units are calories per 24 hours. These are the same calories used to describe the energy content of foods.)

1792 1666 1362 1614 1460 1867 1439

The researchers reported $\bar{x}$ and s for these men.
First find the mean:

$$\begin{aligned}\bar{x} &= \frac{1792 + 1666 + 1362 + 1614 + 1460 + 1867 + 1439}{7} \\ &= \frac{11,200}{7} = 1600 \text{ calories}\end{aligned}$$

Figure 2.3 displays the data as points above the number line, with their mean marked by an asterisk (*). The arrows mark two of the deviations from the mean. These deviations show how spread out the data are about their mean. They are the starting point for calculating the variance and the standard deviation.

Observations x_i	Deviations $x_i - \bar{x}$	Squared deviations $(x_i - \bar{x})^2$
1792	$1792 - 1600 = 192$	$192^2 = 36,864$
1666	$1666 - 1600 = 66$	$66^2 = 4,356$
1362	$1362 - 1600 = -238$	$(-238)^2 = 56,644$
1614	$1614 - 1600 = 14$	$14^2 = 196$
1460	$1460 - 1600 = -140$	$(-140)^2 = 19,600$
1867	$1867 - 1600 = 267$	$267^2 = 71,289$
1439	$1439 - 1600 = -161$	$(-161)^2 = 25,921$
sum = 0		sum = 214,870

The variance is the sum of the squared deviations divided by one less than the number of observations:

$$s^2 = \frac{214,870}{6} = 35,811.67$$

The standard deviation is the square root of the variance:

$$s = \sqrt{35,811.67} = 189.24 \text{ calories}$$

Notice that the “average” in the variance s^2 divides the sum by one fewer than the number of observations, that is, $n - 1$ rather than n . The reason is that the deviations $x_i - \bar{x}$ always sum to exactly 0, so that knowing $n - 1$ of them determines the last one. Only $n - 1$ of the squared deviations can vary freely,

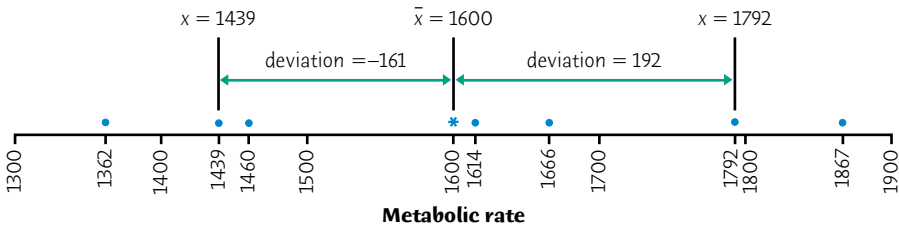


Figure 2.3 Metabolic rates for 7 men, with their mean (*) and the deviations of two observations from the mean.

degrees of freedom

and we average by dividing the total by $n - 1$. The number $n - 1$ is called the **degrees of freedom** of the variance or standard deviation. Many calculators offer a choice between dividing by n and dividing by $n - 1$, so be sure to use $n - 1$.

More important than the details of hand calculation are the properties that determine the usefulness of the standard deviation:

- s measures spread about the mean and should be used only when the mean is chosen as the measure of center.
- $s = 0$ only when there is *no spread*. This happens only when all observations have the same value. Otherwise, $s > 0$. As the observations become more spread out about their mean, s gets larger.
- s has the same units of measurement as the original observations. For example, if you measure metabolic rates in calories, s is also in calories. This is one reason to prefer s to the variance s^2 , which is in squared calories.
- Like the mean $\bar{x}$, s is not resistant. Strong skewness or a few outliers can greatly increase s . For example, the standard deviation of the incomes for our 15 college graduates is 24.9. (Use your calculator to verify this.) If we omit the high outlier, the standard deviation drops to 17.7.

You may rightly feel that the importance of the standard deviation is not yet clear. We will see in the next chapter that the standard deviation is the natural measure of spread for an important class of symmetric distributions, the Normal distributions. The usefulness of many statistical procedures is tied to distributions of particular shapes. This is certainly true of the standard deviation.

Choosing measures of center and spread

How do we choose between the five-number summary and $\bar{x}$ and s to describe the center and spread of a distribution? Because the two sides of a strongly skewed distribution have different spreads, no single number such as s describes the spread well. The five-number summary, with its two quartiles and two extremes, does a better job.

CHOOSING A SUMMARY

The five-number summary is usually better than the mean and standard deviation for describing a skewed distribution or a distribution with strong outliers. Use $\bar{x}$ and s only for reasonably symmetric distributions that are free of outliers.

Do remember that a graph gives the best overall picture of a distribution. Numerical measures of center and spread report specific facts about a distribution, but they do not describe its entire shape. Numerical summaries do not disclose the presence of multiple peaks or gaps, for example. Exercise 2.9 shows how misleading numerical summaries can be. **Always plot your data.**

APPLY YOUR KNOWLEDGE

2.8 Blood phosphate. The level of various substances in the blood influences our health. Here are measurements of the level of phosphate in the blood of a patient, in milligrams of phosphate per deciliter of blood, made on 6 consecutive visits to a clinic:

5.6 5.2 4.6 4.9 5.7 6.4

A graph of only 6 observations gives little information, so we proceed to compute the mean and standard deviation.

- (a) Find the mean from its definition. That is, find the sum of the 6 observations and divide by 6.
- (b) Find the standard deviation from its definition. That is, find the deviations of each observation from the mean, square the deviations, then obtain the variance and the standard deviation. Example 2.6 shows the method.
- (c) Now enter the data into your calculator and use the mean and standard deviation buttons to obtain $\bar{x}$ and s . Do the results agree with your hand calculations?

2.9 $\bar{x}$ and s are not enough. The mean $\bar{x}$ and standard deviation s measure center and spread but are not a complete description of a distribution. Data sets with different shapes can have the same mean and standard deviation. To demonstrate this fact, use your calculator to find $\bar{x}$ and s for these two small data sets. Then make a stemplot of each and comment on the shape of each distribution.

Data A	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74
Data B	6.58	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.50

2.10 Choose a summary. The shape of a distribution is a rough guide to whether the mean and standard deviation are a helpful summary of center and spread. For which of these distributions would $\bar{x}$ and s be useful? In each case, give a reason for your decision.

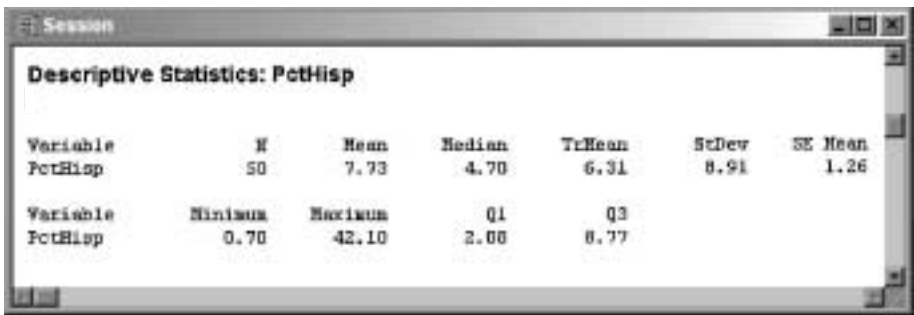
- (a) Percents of Hispanics in the states, Figure 1.3 (page 11).
- (b) Iowa Test scores, Figure 1.4 (page 13).
- (c) Annual costs for Massachusetts colleges, Figure 1.5 (page 14).

Using technology

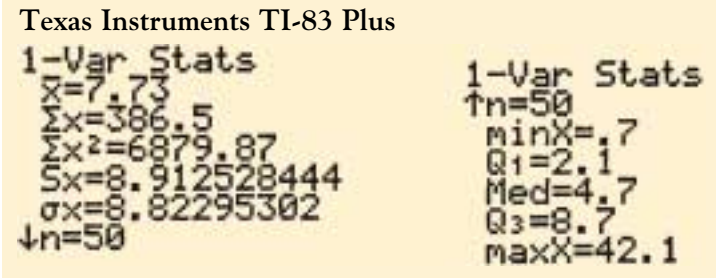
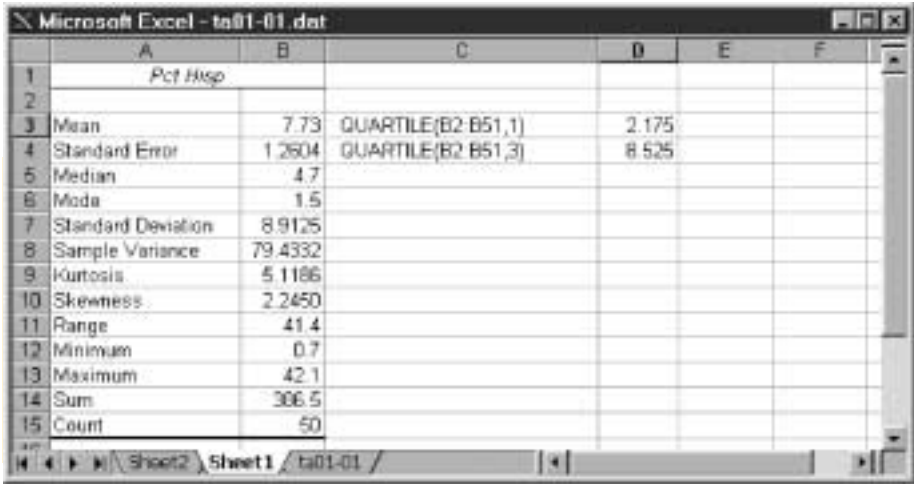
Although a “two-variables statistics” calculator will do the basic calculations we need, more elaborate tools are very helpful. Figure 2.4 displays output describing the percents of Hispanics in the states (Table 1.1, page 10) from three types of tools. Minitab is a statistical software package, Microsoft Excel is spreadsheet software that includes statistical functions, and the Texas

Figure 2.4 Describing the data on the percents of Hispanic residents in the states: output from statistical software, spreadsheet software, and a graphing calculator.

Minitab



Microsoft Excel



Instruments TI-83 Plus is a graphing calculator with statistical functions. All three tools will make statistical graphs, which is very handy. They also do many more statistical calculations than a basic calculator.

Figure 2.4 shows the summary statistics produced by each tool. We are interested in the number of observations n , the five-number summary, and $\bar{x}$ and s . All three displays give this information, but they include other measures that we don't need. This is typical of software—learn to look for what you want and ignore other output. Excel's "Descriptive Statistics" menu item doesn't give the quartiles. We used Excel's separate quartile function to get Q_1 and Q_3 .

EXAMPLE 2.7

What is the first quartile?

What is the first quartile of the percents of Hispanics? Arrange the 50 observations in order:

0.7	0.7	0.9	1.2	1.3	1.4	1.5	1.5	1.7	1.9
2.0	2.0	2.1	2.4	2.4	2.8	2.8	2.9	3.2	3.3
3.5	3.6	4.1	4.3	4.7	4.7	4.8	5.2	5.3	5.5
6.4	6.8	7.0	7.2	7.2	7.9	8.0	8.7	9.0	9.4
10.7	13.3	15.1	16.8	17.1	19.7	25.3	32.0	32.4	42.1

Applying our rules, the median is midway between the 25th and 26th observations in the list, the two bold **4.7**s. All the outputs agree that $M = 4.7$.

The first quartile by our rule is the median of the first 25; that is, it is the 13th in the ordered list, $Q_1 = 2.1$. Only the TI-83 gives this answer. Minitab and Excel don't give 2.1, and they differ from each other. The differences are small: the three results are 2.08, 2.1, and 2.175. That's a reminder that there are several rules for exactly determining the quartiles. Our rule is simplest for hand computation. Results from the various rules are always close to each other, so to describe data you should just use the answer your technology gives you.

Chapter 2 SUMMARY

A numerical summary of a distribution should report its **center** and its **spread** or **variability**.

The **mean** $\bar{x}$ and the **median** M describe the center of a distribution in different ways. The mean is the arithmetic average of the observations, and the median is the midpoint of the values.

When you use the median to indicate the center of the distribution, describe its spread by giving the **quartiles**. The **first quartile** Q_1 has one-fourth of the observations below it, and the **third quartile** Q_3 has three-fourths of the observations below it.

The **five-number summary** consisting of the median, the quartiles, and the high and low extremes provides a quick overall description of a distribution. The median describes the center, and the quartiles and extremes show the spread.

Boxplots based on the five-number summary are useful for comparing several distributions. The box spans the quartiles and shows the spread of the central half of the distribution. The median is marked within the box. Lines extend from the box to the extremes and show the full spread of the data.

The **variance** s^2 and especially its square root, the **standard deviation** s , are common measures of spread about the mean as center. The standard deviation s is zero when there is no spread and gets larger as the spread increases.

A **resistant measure** of any aspect of a distribution is relatively unaffected by changes in the numerical value of a small proportion of the total number of observations, no matter how large these changes are. The median and quartiles are resistant, but the mean and the standard deviation are not.

The mean and standard deviation are good descriptions for symmetric distributions without outliers. They are most useful for the Normal distributions introduced in the next chapter. The five-number summary is a better description for skewed distributions.

Numerical summaries do not fully describe the shape of a distribution. Always plot your data.

Chapter 2 EXERCISES

2.11 Incomes of college grads. The Census Bureau reports that the mean and median income of people who worked full-time in 2001 and had at least a bachelor's degree were \$53,054 and \$72,674. Which of these numbers is the mean and which is the median? Explain your reasoning.

2.12 Internet access. How much do users pay for Internet access? Here are the monthly fees (in dollars) paid by a random sample of 50 users of commercial Internet service providers in August 2000:⁴

20	40	22	22	21	21	20	10	20	20
20	13	18	50	20	18	15	8	22	25
22	10	20	22	22	21	15	23	30	12
9	20	40	22	29	19	15	20	20	20
20	15	19	21	14	22	21	35	20	22

- (a) Make a stemplot of these data. Briefly describe the pattern you see. About how much do you think America Online and its larger competitors were charging in August 2000? Are there any outliers?
- (b) To report a quick summary of how much people pay for Internet service, do you prefer $\bar{x}$ and s or the five-number summary? Why? Calculate your preferred summary.

2.13 Where are the doctors? Table 1.6 (page 29) gives the number of medical doctors per 100,000 people in each state. Exercise 1.29 (page 28) asked you to plot the data. The distribution is right-skewed with several high outliers.

- (a) Do you expect the mean to be greater than the median, about equal to the median, or less than the median? Why? Calculate $\bar{x}$ and M and verify your expectation.
- (b) The District of Columbia is a high outlier at 758 M.D.'s per 100,000 residents. If you remove D.C. because it is a city rather than a state, do you expect $\bar{x}$ or M to change more? Why? Omitting D.C., calculate both measures for the 50 states and verify your expectation.

2.14 McGwire versus Ruth. Exercises 1.25 and 1.26 (page 27) give the numbers of home runs hit each season by Babe Ruth and Mark McGwire. Find the five-number summaries and make side-by-side boxplots to compare these two home run hitters. What do your plots show? Why are the boxplots less informative than the back-to-back stemplot you made in Exercise 1.26 (page 27)?

2.15 Students' attitudes. Here are the scores of 18 first-year college women on the Survey of Study Habits and Attitudes (SSHA):

154 109 137 115 152 140 154 178 101
103 126 126 137 165 165 129 200 148

- (a) Find the mean score from the formula for the mean. Then enter the data into your calculator and use the calculator's $\bar{x}$ button to obtain the mean. Verify that you get the same result.
- (b) A stemplot (Exercise 1.8, page 17) suggests that the score 200 is an outlier. Use your calculator to find the mean for the 17 observations that remain when you drop the outlier. How does the outlier change the mean?

2.16 How much fruit do adolescent girls eat? Figure 1.10 (page 23) is a histogram of the number of servings of fruit per day claimed by 74 seventeen-year-old girls. With a little care, you can find the median and the quartiles from the histogram. What are these numbers? How did you find them?

2.17 Students' attitudes. In Exercise 2.15 you found the mean of the SSHA scores of 18 first-year college women. Now find the median of these scores. Is the median smaller or larger than the mean? Explain why this is so.

2.18 Weight of newborns. Exercise 1.18 (page 24) gives the weight class of all babies born in the United States in 1999.

- (a) How many babies were there? What are the positions of the median and quartiles in the ordered list of all birth weights?
- (b) In which weight classes do the median and quartiles fall?

2.19 Does breast-feeding weaken bones? Breast-feeding mothers secrete calcium into their milk. Some of the calcium may come from their bones, so mothers may lose bone mineral. Researchers compared 47 breast-feeding women with 22 women of similar age who were neither pregnant nor lactating. They measured the percent change in the mineral content of the women's spines over three months. Here are the data:⁵



(Tony Anderson/Taxi/Getty Images)

Breast-feeding women						Other women					
-4.7	-2.5	-4.9	-2.7	-0.8	-5.3	2.4	0.0	0.9	-0.2	1.0	1.7
-8.3	-2.1	-6.8	-4.3	2.2	-7.8	2.9	-0.6	1.1	-0.1	-0.4	0.3
-3.1	-1.0	-6.5	-1.8	-5.2	-5.7	1.2	-1.6	-0.1	-1.5	0.7	-0.4
-7.0	-2.2	-6.5	-1.0	-3.0	-3.6	2.2	-0.4	-2.2	-0.1		
-5.2	-2.0	-2.1	-5.6	-4.4	-3.3						
-4.0	-4.9	-4.7	-3.8	-5.9	-2.5						
-0.3	-6.2	-6.8	1.7	0.3	-2.3						
0.4	-5.3	0.2	-2.2	-5.1							

Compare the two distributions using both graphs and numerical summaries. Do the data show distinctly greater bone mineral loss among the breast-feeding women?

2.20 The density of the earth. In 1798 the English scientist Henry Cavendish measured the density of the earth with great care. It is common practice to repeat careful measurements several times and use the mean as the final result. Cavendish repeated his work 29 times. Here are his results (the data give the density of the earth as a multiple of the density of water):⁶

5.50 5.61 4.88 5.07 5.26 5.55 5.36 5.29 5.58 5.65
5.57 5.53 5.62 5.29 5.44 5.34 5.79 5.10 5.27 5.39
5.42 5.47 5.63 5.34 5.46 5.30 5.75 5.68 5.85

Present these measurements with a graph of your choice. Scientists usually give the mean and standard deviation to summarize a set of measurements. Does the shape of this distribution suggest that $\bar{x}$ and s are adequate summaries? Calculate $\bar{x}$ and s .

2.21 California counties. You are planning a sample survey of households in California. You decide to select households separately within each county and to choose more households from the more populous counties. To aid in the planning, Table 2.3 gives the populations of California counties from the 2000 census. Examine the distribution



(Claver Carroll/Photolibrary/
PictureQuest)

TABLE 2.3 Population of California counties, 2000 census

County	Population	County	Population	County	Population
Alameda	1,443,741	Marin	247,289	San Mateo	707,161
Alpine	1,208	Mariposa	17,130	Santa Barbara	399,347
Amador	35,100	Mendocino	86,265	Santa Clara	1,682,585
Butte	203,171	Merced	210,554	Santa Cruz	255,602
Calaveras	40,554	Modoc	9,449	Shasta	163,256
Colusa	18,804	Mono	12,853	Sierra	3,555
Contra Costa	948,816	Monterey	401,762	Siskiyou	44,301
Del Norte	27,507	Napa	124,279	Solano	394,542
El Dorado	156,299	Nevada	92,033	Sonoma	458,614
Fresno	799,407	Orange	2,846,289	Stanislaus	446,997
Glenn	26,453	Placer	248,399	Sutter	78,930
Humboldt	126,518	Plumas	20,824	Tehama	56,039
Imperial	142,361	Riverside	1,545,387	Trinity	13,022
Inyo	17,945	Sacramento	1,223,499	Tulare	368,021
Kern	661,645	San Benito	53,234	Tuolumne	54,501
Kings	129,461	San Bernardino	1,709,434	Ventura	753,197
Lake	58,309	San Diego	2,813,833	Yolo	168,660
Lassen	33,828	San Francisco	776,733	Yuba	60,219
Los Angeles	9,519,338	San Joaquin	563,598		
Madera	123,109	San Luis Obispo	246,681		

TABLE 2.4 Positions and weights (pounds) for a major college football team

QB	200	QB	209	QB	190	QB	201	RB	210	RB	224	RB	196
RB	218	RB	229	RB	236	OL	281	OL	286	OL	320	OL	369
OL	298	OL	276	OL	293	OL	292	OL	285	OL	286	OL	265
OL	314	OL	318	OL	334	OL	276	OL	300	OL	290	WR	227
WR	178	WR	180	WR	193	WR	190	WR	163	WR	185	WR	200
TE	235	TE	233	TE	225	TE	253	TE	275	TE	200	KP	189
KP	214	KP	185	KP	204	DB	189	DB	173	DB	186	DB	220
DB	170	DB	166	DB	185	DB	194	DB	175	DB	194	DB	193
DB	179	LB	247	LB	208	LB	227	LB	219	LB	219	LB	208
LB	225	LB	203	DL	240	DL	261	DL	302	DL	214	DL	225
DL	250	DL	350	DL	230	DL	271	DL	294	DL	202	DL	297
DL	271	DL	260	DL	257	DL	262						

of county populations both graphically and numerically, using whatever tools are most suitable. Write a brief description of the main features of this distribution. Sample surveys often select households from all of the most populous counties but from only some of the less populous. How would you divide California counties into three groups according to population, with the intent of including all of the first group, half of the second, and a smaller fraction of the third in your survey?

2.22 A football team. The University of Miami Hurricanes have been among the more successful teams in college football. Table 2.4 gives the weights in pounds and positions of the players on the 2002 team.⁷ The positions are quarterback (QB), running back (RB), offensive line (OL), wide receiver (WR), tight end (TE), kicker/punter (KP), defensive back (DB), linebacker (LB), and defensive line (DL).

- (a) Make side-by-side boxplots of the weights for running backs, wide receivers, offensive linemen, defensive linemen, linebackers, and defensive backs.
- (b) Briefly compare the weight distributions. Which position has the heaviest players overall? Which has the lightest?
- (c) Are any individual players outliers within their position?



(AP/Wide World Photos)

2.23 Guinea pig survival times. Here are the survival times in days of 72 guinea pigs after they were injected with infectious bacteria in a medical experiment.⁸ Survival times, whether of machines under stress or cancer patients after treatment, usually have distributions that are skewed to the right.

43	45	53	56	56	57	58	66	67	73	74	79
80	80	81	81	81	82	83	83	84	88	89	91
91	92	92	97	99	99	100	100	101	102	102	102
103	104	107	108	109	113	114	118	121	123	126	128
137	138	139	144	145	147	156	162	174	178	179	184
191	198	211	214	243	249	329	380	403	511	522	598

- (a) Graph the distribution and describe its main features. Does it show the expected right skew?
- (b) Which numerical summary would you choose for these data? Calculate your chosen summary. How does it reflect the skewness of the distribution?

2.24 Never on Sunday: also in Canada? Exercise 1.4 (page 9) gives the number of births in the United States on each day of the week during an entire year. The boxplots in Figure 2.5 are based on more detailed data from Toronto, Canada: the number of births on each of the 365 days in a year, grouped by day of the week.⁹ Based on these plots, give a more detailed description of how births depend on the day of the week.

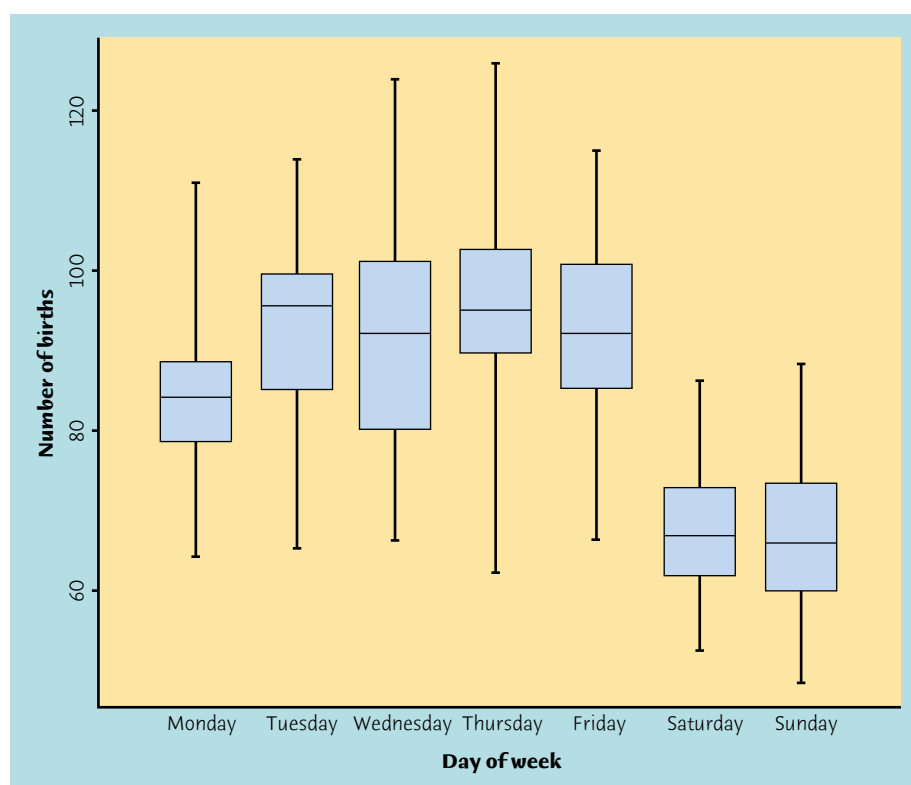


Figure 2.5 Side-by-side boxplots of the distributions of numbers of births in Toronto, Canada, for each day of the week during a year.

2.25 Athletes' salaries. In 2002, the Chicago Cubs failed once again to reach the National League playoffs. Table 2.5 gives the salaries of the players as of opening day of the season. Describe the distribution of salaries both with a graph and with a numerical summary. Then write a brief description of the important features of the distribution.

Identifying suspected outliers. Whether an observation is an outlier is a matter of judgment: does it appear to clearly stand apart from the rest

TABLE 2.5 2002 salaries for the Chicago Cubs baseball team

Player	Salary	Player	Salary	Player	Salary
Sammy Sosa	\$15,000,000	Jason Bere	\$3,400,000	Donovan Osborne	\$450,000
Fred McGriff	\$7,250,000	Jeff Fassero	\$2,800,000	Kyle Farnsworth	\$410,000
Todd Hundley	\$6,500,000	Tom Gordon	\$2,800,000	Robert Machado	\$277,500
Moises Alou	\$6,000,000	Matt Clement	\$2,500,000	Roosevelt Brown	\$255,000
Jon Lieber	\$5,833,333	Joe Girardi	\$1,500,000	Augie Ojeda	\$240,000
Alex Gonzalez	\$4,250,000	Jesus Sanchez	\$1,425,000	Joe Borowski	\$235,000
Kerry Wood	\$3,695,000	Delino DeShields	\$1,250,000	Corey Patterson	\$227,500
Antonio Alfonseca	\$3,550,000	Chris Stynes	\$1,250,000	Mark Bellhorn	\$224,000
Bill Mueller	\$3,450,000	Darren Lewis	\$500,000	Juan Cruz	\$215,000
				Will Ohman	\$203,500

of the distribution? When large volumes of data are scanned automatically, however, we need a rule to pick out suspected outliers. The most common rule is the **1.5IQR rule**. The **interquartile range** is the distance between the two quartiles, $IQR = Q_3 - Q_1$. This is the spread of the middle half of the data. A point is a suspected outlier if it lies more than 1.5IQR below the first quartile Q_1 or above the third quartile Q_3 . Exercises 2.26 to 2.28 ask you to apply this rule.

1.5IQR rule
interquartile range

- 2.26 A high income.** In Example 2.1, we noted the influence of one high income of \$110,000 among the incomes of a sample of 15 college graduates. Does the 1.5IQR rule identify this income as a suspected outlier?
- 2.27 Athletes' salaries.** Which members of the Chicago Cubs (see Table 2.5) have salaries that are suspected outliers by the 1.5IQR rule?
- 2.28 Gas guzzlers?** The BMW Z8, Ferrari, and Lamborghini have notably low gas mileage among the two-seater cars in Table 1.2 (page 12). Find the five-number summary of the city gas mileages (omitting the Honda Insight). Are any of these cars suspected low outliers by the 1.5IQR rule?
- 2.29 A standard deviation contest.** This is a standard deviation contest. You must choose four numbers from the whole numbers 0 to 10, with repeats allowed.
- (a) Choose four numbers that have the smallest possible standard deviation.
 - (b) Choose four numbers that have the largest possible standard deviation.
 - (c) Is more than one choice possible in either (a) or (b)? Explain.
- 2.30 Test your technology.** This exercise requires a calculator with a standard deviation button or statistical software on a computer.



(Kim Sayer/Corbis)

The observations

10,001 10,002 10,003

have mean $\bar{x} = 10,002$ and standard deviation $s = 1$. Adding a 0 in the center of each number, the next set becomes

100,001 100,002 100,003

The standard deviation remains $s = 1$ as more 0s are added. Use your calculator or software to find the standard deviation of these numbers, adding extra 0s until you get an incorrect answer. How soon did you go wrong? This demonstrates that calculators and software cannot handle an arbitrary number of digits correctly.

- 2.31 You create the data.** Create a set of 5 positive numbers (repeats allowed) that have median 10 and mean 7. What thought process did you use to create your numbers?
- 2.32 You create the data.** Give an example of a small set of data for which the mean is larger than the third quartile.

Chapter 2 MEDIA EXERCISES



- 2.33 Comparing incomes.** Want data on incomes? Go to the Census Bureau Web site, www.census.gov, click on “Income,” and open the latest issue of the annual publication *Money Income in the United States*. The tables in this publication contain lots of detail. One set of tables compares the distributions of incomes among households by ethnicity (white, black, Asian and Pacific Islander, and Hispanic origin). What is the median household income for each group? How closely can you find the quartiles from the information given? Make partial boxplots—just the central box—to compare the four distributions. What do the data show?



- 2.34 Internet access.** You can use the *One-Variable Statistical Calculator* applet in place of a calculator or software to do both calculations ($\bar{x}$ and s and the five-number summary) and graphs (histograms and stemplots). The applet is more convenient than most calculators. It is less convenient than good software because, depending on your browser, it may be difficult to enter new data sets in one operation and to print output. Exercise 2.12 (page 48) gives the monthly fees paid for Internet access by a random sample of customers of commercial Internet service providers. This data set is one of those stored in the applet. Use the applet to do Exercise 2.12.



- 2.35 Mean = median?** The *Mean and Median* applet allows you to place observations on a line and see their mean and median visually. Place two observations on the line. Why does only one arrow appear?

2.36 Making resistance visible. In the *Mean and Median* applet, place three observations on the line by clicking below it: two close together near the center of the line and one somewhat to the right of these two.



- (a) Pull the single rightmost observation out to the right. (Place the cursor on the point, hold down a mouse button, and drag the point.) How does the mean behave? How does the median behave? Explain briefly why each measure acts as it does.
- (b) Now drag the single rightmost point to the left as far as you can. What happens to the mean? What happens to the median as you drag this point past the other two (watch carefully)?

2.37 Behavior of the median. Place five observations on the line in the *Mean and Median* applet by clicking below it.



- (a) Add one additional observation *without changing the median*. Where is your new point?
- (b) Use the applet to convince yourself that when you add yet another observation (there are now seven in all), the median does not change no matter where you put the seventh point. Explain why this must be true.

2.38 Eruptions of Old Faithful. In Exercise 1.37 (page 31) you used a graph to describe how long eruptions of famous Old Faithful geyser last. Data for 299 eruptions appear in the EESEE story “Is Old Faithful Faithful?” Now you will continue your work.

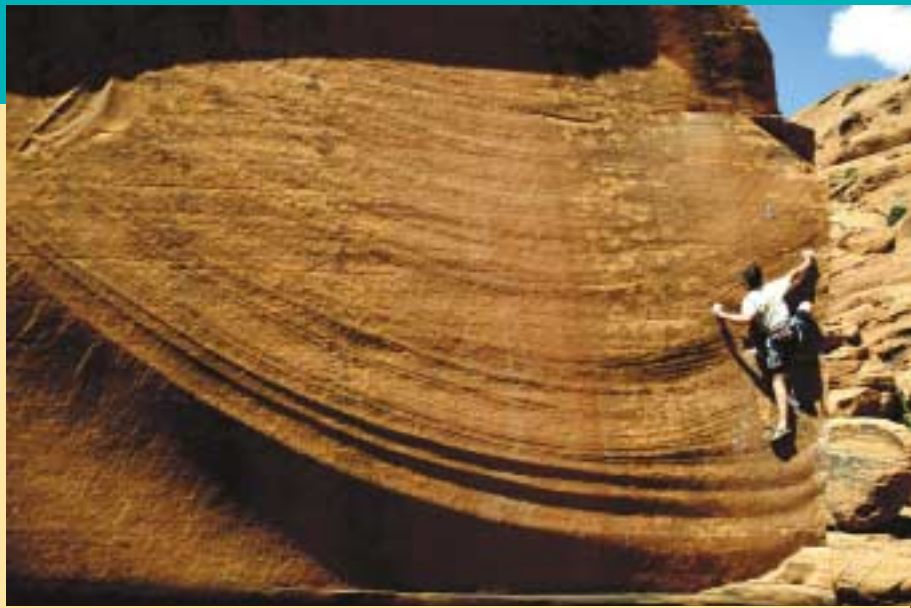


- (a) The shape of the distribution fits the fact that Old Faithful has two reservoirs of water, which can erupt either separately or together. Explain why the shape matches this fact. Also explain why neither $\bar{x}$ and s nor the five-number summary is a good description of this distribution.
- (b) Use your judgment, based on the histogram, to divide the eruptions into two clusters. What is the shape of the distribution of eruption lengths in each cluster? Give numerical descriptions of both clusters and compare the two.

2.39 Truth in calories. The EESEE story “Counting Calories” presents data from a study showing that foods advertised as “low calorie” and the like often contain more calories than the label states. One question was whether the degree of understatement of calories is the same for national brands, regional brands, and locally prepared foods. The EESEE story has data on the percent difference between true and label calories for many foods of each type. The differences are greater than zero when the true calories are higher than the label claims. Use both graphs and numerical summaries to compare the three types of food. What are your conclusions?



CHAPTER 3



(Mike Powell/Allsport Concepts/Getty Images)

In this chapter we cover...

Density curves

The median and mean of a density curve

Normal distributions

The 68–95–99.7 rule

The standard Normal distribution

Normal distribution calculations

Finding a value given a proportion

The Normal Distributions

We now have a kit of graphical and numerical tools for describing distributions. What is more, we have a clear strategy for exploring data on a single quantitative variable:

1. Always plot your data: make a graph, usually a histogram or a stemplot.
2. Look for the overall pattern (shape, center, spread) and for striking deviations such as outliers.
3. Calculate a numerical summary to briefly describe center and spread.

Here is one more step to add to this strategy:

4. Sometimes the overall pattern of a large number of observations is so regular that we can describe it by a smooth curve.

Density curves

Figure 3.1 is a histogram of the scores of all 947 seventh-grade students in Gary, Indiana, on the vocabulary part of the Iowa Test of Basic Skills.¹ Scores of many students on this national test have a quite regular distribution. The histogram is symmetric, and both tails fall off quite smoothly from a single center peak. There are no large gaps or obvious outliers. The smooth curve drawn through the tops of the histogram bars in Figure 3.1 is a good description of the overall pattern of the data. The curve is a **mathematical model** for the distribution. A mathematical model is an idealized description. It gives a compact picture

mathematical model

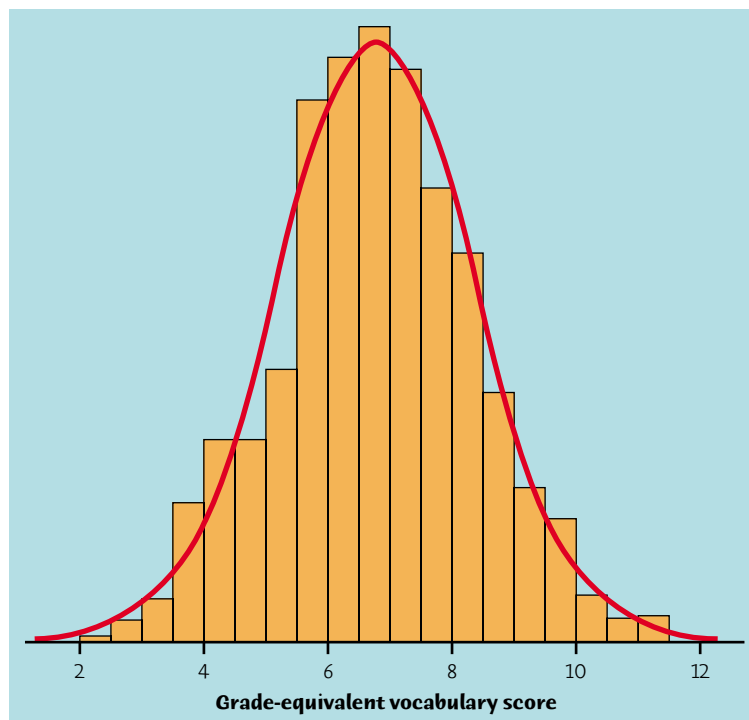


Figure 3.1 Histogram of the vocabulary scores of all seventh-grade students in Gary, Indiana. The smooth curve shows the overall shape of the distribution.

of the overall pattern of the data but ignores minor irregularities as well as any outliers.

We will see that it is easier to work with the smooth curve in Figure 3.1 than with the histogram. The reason is that the histogram depends on our choice of classes, while with a little care we can use a curve that does not depend on any choices we make. Here's how we do it.

EXAMPLE 3.1 From histogram to density curve

Our eyes respond to the *areas* of the bars in a histogram. The bar areas represent proportions of the observations. Figure 3.2(a) is a copy of Figure 3.1 with the leftmost bars shaded. The area of the shaded bars in Figure 3.2(a) represents the students with vocabulary scores 6.0 or lower. There are 287 such students, who make up the proportion $287/947 = 0.303$ of all Gary seventh graders.

Now concentrate on the curve drawn through the bars. In Figure 3.2(b), the area under the curve to the left of 6.0 is shaded. Adjust the scale of the graph so that *the total area under the curve is exactly 1*. This area represents the proportion 1, that is, all the observations. Areas under the curve then represent proportions of the observations. The curve is now a *density curve*. The shaded area under the density curve in Figure 3.2(b) represents the proportion of students with score 6.0 or lower. This area is 0.293, only 0.010 away from the histogram result. You can see that areas under the density curve give quite good approximations of areas given by the histogram.

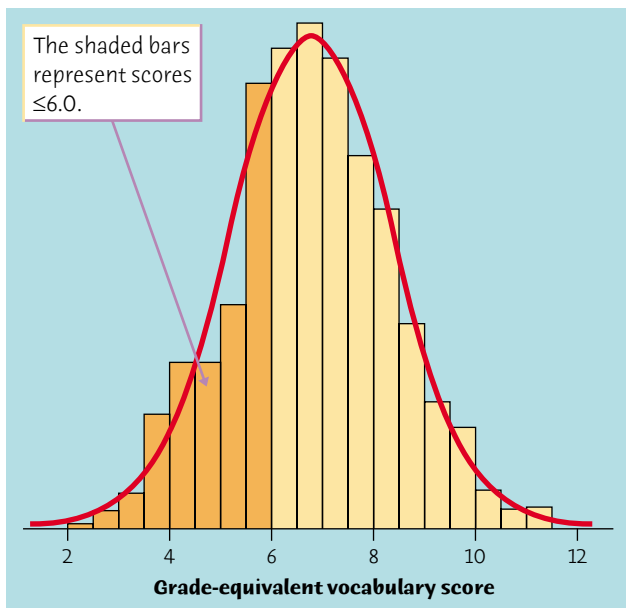


Figure 3.2(a) The proportion of scores less than or equal to 6.0 from the histogram is 0.303.

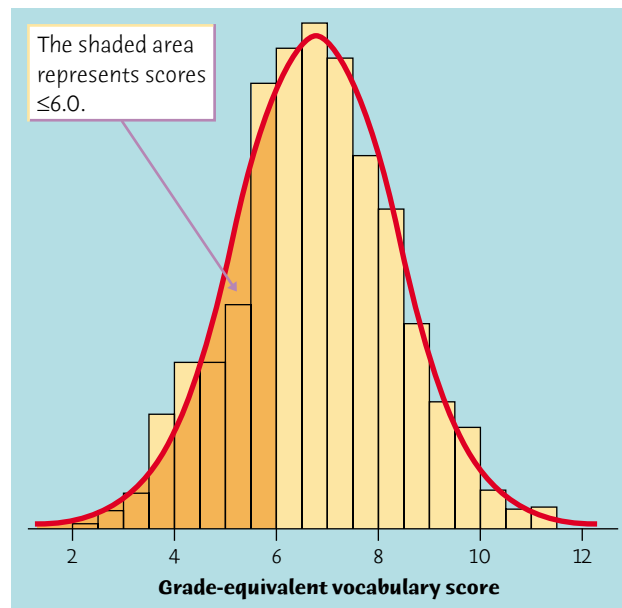


Figure 3.2(b) The proportion of scores less than or equal to 6.0 from the density curve is 0.293.

DENSITY CURVE

A **density curve** is a curve that

- is always on or above the horizontal axis, and
- has area exactly 1 underneath it.

A density curve describes the overall pattern of a distribution. The area under the curve and above any range of values is the proportion of all observations that fall in that range.

Normal curve

The density curve in Figures 3.1 and 3.2 is a **Normal curve**. Density curves, like distributions, come in many shapes. Figure 3.3 shows two density curves: a symmetric Normal density curve and a right-skewed curve. A density curve of the appropriate shape is often an adequate description of the overall pattern of a distribution. Outliers, which are deviations from the overall pattern, are not described by the curve. Of course, no set of real data is exactly described by a density curve. The curve is an approximation that is easy to use and accurate enough for practical use.

APPLY YOUR KNOWLEDGE

3.1 Sketch density curves. Sketch density curves that might describe distributions with the following shapes:

- (a) Symmetric, but with two peaks (that is, two strong clusters of observations).
- (b) Single peak and skewed to the left.

The median and mean of a density curve

Our measures of center and spread apply to density curves as well as to actual sets of observations. The median and quartiles are easy. Areas under a density curve represent proportions of the total number of observations. The median is the point with half the observations on either side. So **the median of a density curve is the equal-areas point**, the point with half the area under the curve to its left and the remaining half of the area to its right. The quartiles divide the area under the curve into quarters. One-fourth of the area under the curve is to the left of the first quartile, and three-fourths of the area is to the left of the third quartile. You can roughly locate the median and quartiles of any density curve by eye by dividing the area under the curve into four equal parts.

Because density curves are idealized patterns, a symmetric density curve is exactly symmetric. The median of a symmetric density curve is therefore at its center. Figure 3.3(a) shows the median of a symmetric curve. It isn't so easy to spot the equal-areas point on a skewed curve. There are mathematical ways of finding the median for any density curve. We did that to mark the median on the skewed curve in Figure 3.3(b).

What about the mean? The mean of a set of observations is their arithmetic average. If we think of the observations as weights strung out along a thin rod, the mean is the point at which the rod would balance. This fact is also true of density curves. **The mean is the point at which the curve would balance if made of solid material.** Figure 3.4 illustrates this fact about the mean. A symmetric curve balances at its center because the two sides are identical. **The mean and median of a symmetric density curve are equal**, as in Figure 3.3(a). We know that the mean of a skewed distribution is pulled toward the long tail.

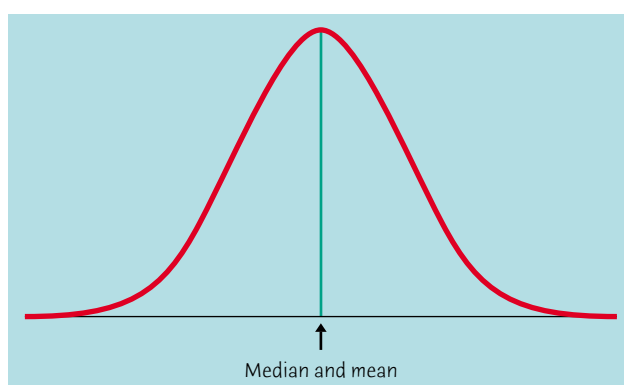


Figure 3.3(a) The median and mean of a symmetric density curve.

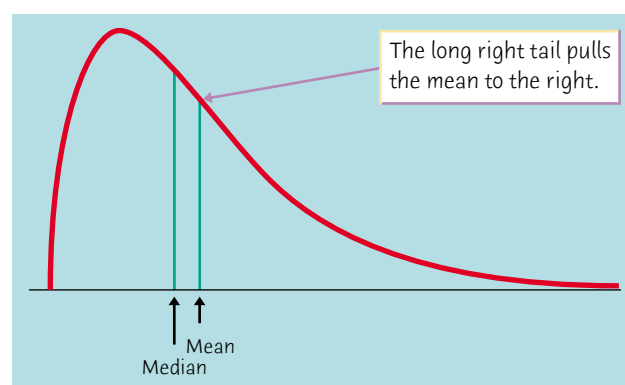


Figure 3.3(b) The median and mean of a right-skewed density curve.

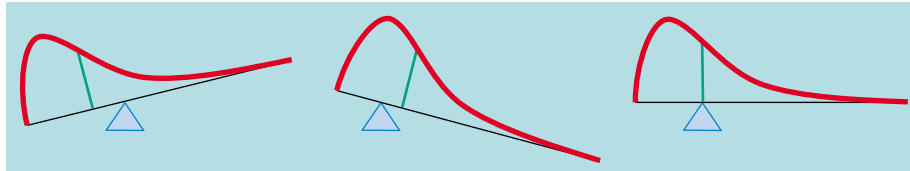


Figure 3.4 The mean is the balance point of a density curve.

Figure 3.3(b) shows how the mean of a skewed density curve is pulled toward the long tail more than is the median. It's hard to locate the balance point by eye on a skewed curve. There are mathematical ways of calculating the mean for any density curve, so we are able to mark the mean as well as the median in Figure 3.3(b).

MEDIAN AND MEAN OF A DENSITY CURVE

The **median** of a density curve is the equal-areas point, the point that divides the area under the curve in half.

The **mean** of a density curve is the balance point, at which the curve would balance if made of solid material.

The median and mean are the same for a symmetric density curve. They both lie at the center of the curve. The mean of a skewed curve is pulled away from the median in the direction of the long tail.

We can roughly locate the mean, median, and quartiles of any density curve by eye. This is not true of the standard deviation. When necessary, we can once again call on more advanced mathematics to learn the value of the standard deviation. The study of mathematical methods for doing calculations with density curves is part of theoretical statistics. Though we are concentrating on statistical practice, we often make use of the results of mathematical study.

Because a density curve is an idealized description of the distribution of data, we need to distinguish between the mean and standard deviation of the density curve and the mean $\bar{x}$ and standard deviation s computed from the actual observations. The usual notation for the mean of an idealized distribution is μ (the Greek letter mu). We write the standard deviation of a density curve as σ (the Greek letter sigma).

mean μ
standard deviation σ

APPLY YOUR KNOWLEDGE

3.2 A uniform distribution. Figure 3.5 displays the density curve of a *uniform distribution*. The curve takes the constant value 1 over the interval from 0 to 1 and is zero outside that range of values. This means that data described by this distribution take values that are uniformly spread between 0 and 1. Use areas under this density curve to answer the following questions.

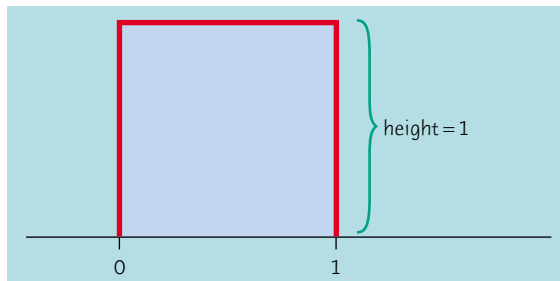


Figure 3.5 The density curve of a uniform distribution, for Exercise 3.2.

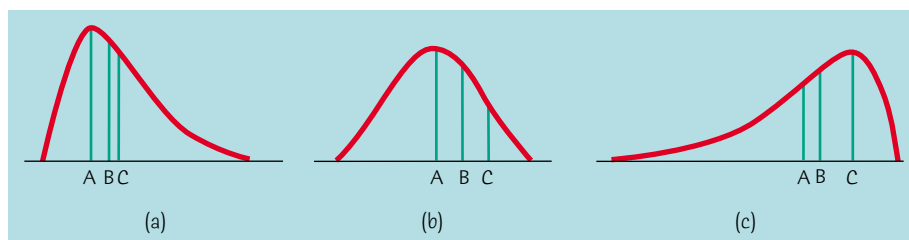


Figure 3.6 Three density curves, for Exercise 3.4.

- (a) Why is the total area under this curve equal to 1?
- (b) What percent of the observations lie above 0.8?
- (c) What percent of the observations lie below 0.6?
- (d) What percent of the observations lie between 0.25 and 0.75?

3.3 Mean and median. What is the mean μ of the density curve pictured in Figure 3.5? What is the median?

3.4 Mean and median. Figure 3.6 displays three density curves, each with three points marked on them. At which of these points on each curve do the mean and the median fall?

Normal distributions

One particularly important class of density curves has already appeared in Figures 3.1 and 3.3(a). These density curves are symmetric, single-peaked, and bell-shaped. They are called *Normal curves*, and they describe **Normal distributions**. Normal distributions play a large role in statistics, but they are rather special and not at all “normal” in the sense of being average or natural. We capitalize Normal to remind you that these curves are special. All Normal distributions have the same overall shape. The exact density curve for a particular Normal distribution is described by giving its mean μ and its standard deviation σ . The mean is located at the center of the symmetric curve and is the same as the median. Changing μ without changing σ moves the Normal curve along the horizontal axis without changing its spread. The standard deviation σ controls

Normal distributions

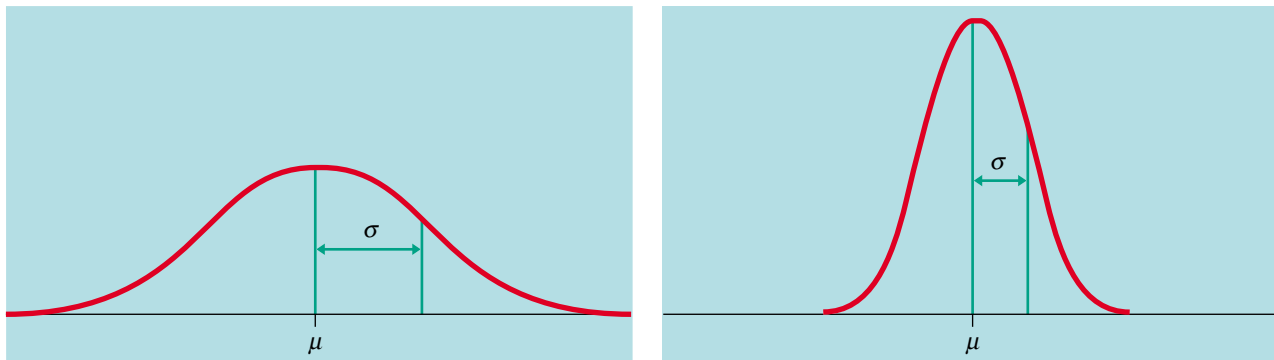


Figure 3.7 Two Normal curves, showing the mean μ and standard deviation σ .

the spread of a Normal curve. Figure 3.7 shows two Normal curves with different values of σ . The curve with the larger standard deviation is more spread out.

The standard deviation σ is the natural measure of spread for Normal distributions. Not only do μ and σ completely determine the shape of a Normal curve, but we can locate σ by eye on the curve. Here's how. Imagine that you are skiing down a mountain that has the shape of a Normal curve. At first, you descend at an ever-steeper angle as you go out from the peak:



Fortunately, before you find yourself going straight down, the slope begins to grow flatter rather than steeper as you go out and down:



The points at which this change of curvature takes place are located at distance σ on either side of the mean μ . You can feel the change as you run a pencil along a Normal curve, and so find the standard deviation. Remember that μ and σ alone do not specify the shape of most distributions, and that the shape of density curves in general does not reveal σ . These are special properties of Normal distributions.

Why are the Normal distributions important in statistics? Here are three reasons. First, Normal distributions are good descriptions for some distributions of *real data*. Distributions that are often close to Normal include scores on tests taken by many people (such as SAT exams and many psychological tests), repeated careful measurements of the same quantity, and characteristics of biological populations (such as lengths of crickets and yields of corn). Second, Normal distributions are good approximations to the results of many kinds of *chance outcomes*, such as tossing a coin many times. Third, and most important, we will see that many *statistical inference* procedures based on Normal distributions work well for other roughly symmetric distributions. However, many sets of data do not follow a Normal distribution. Most income distributions, for

example, are skewed to the right and so are not Normal. Non-Normal data, like nonnormal people, not only are common but are sometimes more interesting than their Normal counterparts.

The 68–95–99.7 rule

Although there are many Normal curves, they all have common properties. In particular, all Normal distributions obey the following rule.

THE 68–95–99.7 RULE

In the Normal distribution with mean μ and standard deviation σ :

- 68% of the observations fall within σ of the mean μ .
- 95% of the observations fall within 2σ of μ .
- 99.7% of the observations fall within 3σ of μ .

Figure 3.8 illustrates the 68–95–99.7 rule. By remembering these three numbers, you can think about Normal distributions without constantly making detailed calculations.

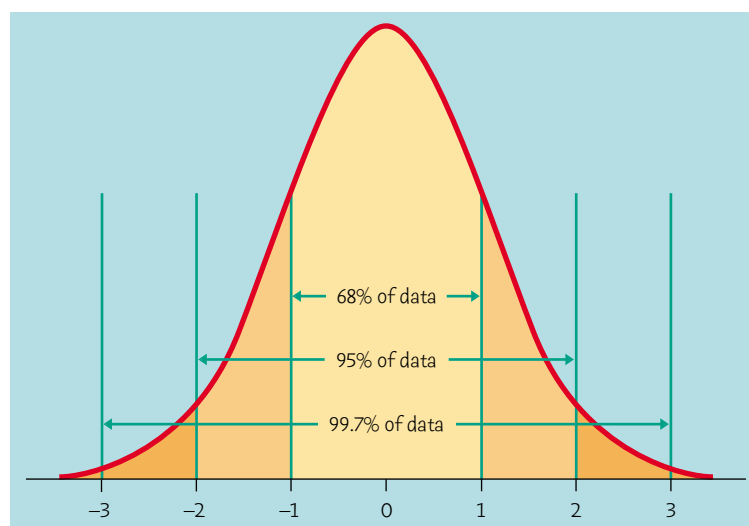


Figure 3.8 The 68–95–99.7 rule for Normal distributions.

EXAMPLE 3.2 Using the 68–95–99.7 rule: SAT scores

The distribution of the scores of the more than 1.3 million high school seniors who took the SAT Reasoning (SAT I) verbal exam in 2002 is close to Normal with mean 504 and standard deviation 111.² Figure 3.9 applies the 68–95–99.7 rule to SAT scores.

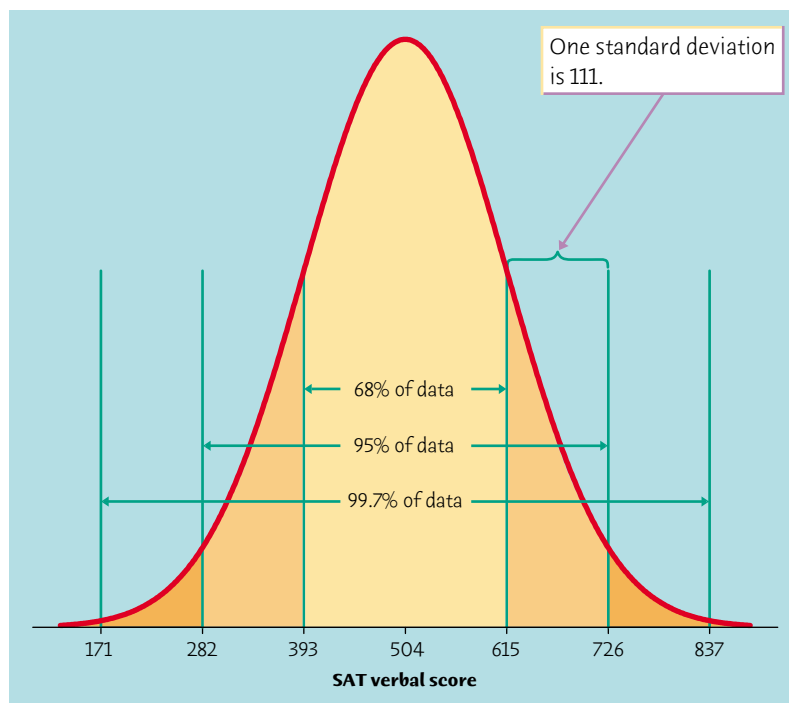


Figure 3.9 The 68–95–99.7 rule applied to the distribution of SAT verbal scores for 2002, with $\mu = 504$ and $\sigma = 111$.

Two standard deviations is 222 for this distribution. The 95 part of the 68–95–99.7 rule says that the middle 95% of SAT verbal scores are between $504 - 222$ and $504 + 222$, that is, between 282 and 726. The other 5% of scores lie outside this range. Because the Normal distributions are symmetric, half of these scores are on the high side. So the highest 2.5% of SAT verbal scores are higher than 726. These facts are exactly true for Normal distributions. They are approximately true for SAT scores, whose distribution is approximately Normal.

The 99.7 part of the 68–95–99.7 rule says that almost all scores (99.7% of them) lie between $\mu - 3\sigma$ and $\mu + 3\sigma$. This range of scores is 171 to 837. In fact, SAT scores are only reported between 200 and 800. It is possible to score higher than 800, but the score will be reported as 800. That is, you don't need a perfect paper to score 800 on the SAT.

Because we will mention Normal distributions often, a short notation is helpful. We abbreviate the Normal distribution with mean μ and standard deviation σ as $N(\mu, \sigma)$. For example, the distribution of SAT verbal scores is $N(504, 111)$.

APPLY YOUR KNOWLEDGE

3.5 Heights of young women. The distribution of heights of women aged 20 to 29 is approximately Normal with mean 64 inches and standard deviation 2.7 inches.³ Draw a Normal curve on which this mean and

standard deviation are correctly located. (Hint: Draw the curve first, locate the points where the curvature changes, then mark the horizontal axis.)

3.6 Heights of young women. The distribution of heights of women aged 20 to 29 is approximately Normal with mean 64 inches and standard deviation 2.7 inches. Use the 68–95–99.7 rule to answer the following questions.

- (a) Between what heights do the middle 95% of young women fall?
- (b) What percent of young women are taller than 61.3 inches?

3.7 Length of pregnancies. The length of human pregnancies from conception to birth varies according to a distribution that is approximately Normal with mean 266 days and standard deviation 16 days. Use the 68–95–99.7 rule to answer the following questions.

- (a) Between what values do the lengths of the middle 95% of all pregnancies fall?
- (b) How short are the shortest 2.5% of all pregnancies?

The standard Normal distribution

As the 68–95–99.7 rule suggests, all Normal distributions share many common properties. In fact, all Normal distributions are the same if we measure in units of size σ about the mean μ as center. Changing to these units is called *standardizing*. To standardize a value, subtract the mean of the distribution and then divide by the standard deviation.

STANDARDIZING AND z -SCORES

If x is an observation from a distribution that has mean μ and standard deviation σ , the **standardized value** of x is

$$z = \frac{x - \mu}{\sigma}$$

A standardized value is often called a **z -score**.

A z -score tells us how many standard deviations the original observation falls away from the mean, and in which direction. Observations larger than the mean are positive when standardized, and observations smaller than the mean are negative.

EXAMPLE 3.3 Standardizing women's heights

The heights of young women are approximately Normal with $\mu = 64$ inches and $\sigma = 2.7$ inches. The standardized height is

$$z = \frac{\text{height} - 64}{2.7}$$



He said, she said.

The height and weight distributions in this chapter come from actual measurements by a government survey. Good thing that is. When *asked* their weight, almost all women say they weigh less than they really do. Heavier men also underreport their weight—but lighter men claim to weigh more than the scale shows. We leave you to ponder the psychology of the two sexes. Just remember that “say so” is no replacement for measuring.

A woman's standardized height is the number of standard deviations by which her height differs from the mean height of all young women. A woman 70 inches tall, for example, has standardized height

$$z = \frac{70 - 64}{2.7} = 2.22$$

or 2.22 standard deviations above the mean. Similarly, a woman 5 feet (60 inches) tall has standardized height

$$z = \frac{60 - 64}{2.7} = -1.48$$

or 1.48 standard deviations less than the mean height.

If the variable we standardize has a Normal distribution, standardizing does more than give a common scale. It makes all Normal distributions into a single distribution, and this distribution is still Normal. Standardizing a variable that has any Normal distribution produces a new variable that has the *standard Normal distribution*.

STANDARD NORMAL DISTRIBUTION

The **standard Normal distribution** is the Normal distribution $N(0, 1)$ with mean 0 and standard deviation 1.

If a variable x has any Normal distribution $N(\mu, \sigma)$ with mean μ and standard deviation σ , then the standardized variable

$$z = \frac{x - \mu}{\sigma}$$

has the standard Normal distribution.

APPLY YOUR KNOWLEDGE

- 3.8 SAT versus ACT.** Eleanor scores 680 on the mathematics part of the SAT. The distribution of SAT math scores in 2002 was Normal with mean 516 and standard deviation 114. Gerald takes the ACT Assessment mathematics test and scores 27. ACT math scores in 2002 were Normally distributed with mean 20.6 and standard deviation 5.0. Find the standardized scores for both students. Assuming that both tests measure the same kind of ability, who has the higher score?
- 3.9 Men's and women's heights.** The heights of women aged 20 to 29 are approximately Normal with mean 64 inches and standard deviation 2.7 inches. Men the same age have mean height 69.3 inches with standard deviation 2.8 inches. What are the z -scores for a woman 6 feet

tall and a man 6 feet tall? Say in simple language what information the z -scores give that the actual heights do not.

Normal distribution calculations

An area under a density curve is a proportion of the observations in a distribution. Any question about what proportion of observations lie in some range of values can be answered by finding an area under the curve. Because all Normal distributions are the same when we standardize, we can find areas under any Normal curve from a single table, a table that gives areas under the curve for the standard Normal distribution.

EXAMPLE 3.4 Using the standard Normal distribution

What proportion of all young women are less than 70 inches tall? This proportion is the area under the $N(64, 2.7)$ curve to the left of the point 70. Because the standardized height corresponding to 70 inches is

$$z = \frac{x - \mu}{\sigma} = \frac{70 - 64}{2.7} = 2.22$$

this area is the same as the area under the standard Normal curve to the left of the point $z = 2.22$. Figure 3.10(a) shows this area.

Calculators and software often give areas under Normal curves. You can do the work by hand using Table A in the front end covers, which gives areas under the standard Normal curve.

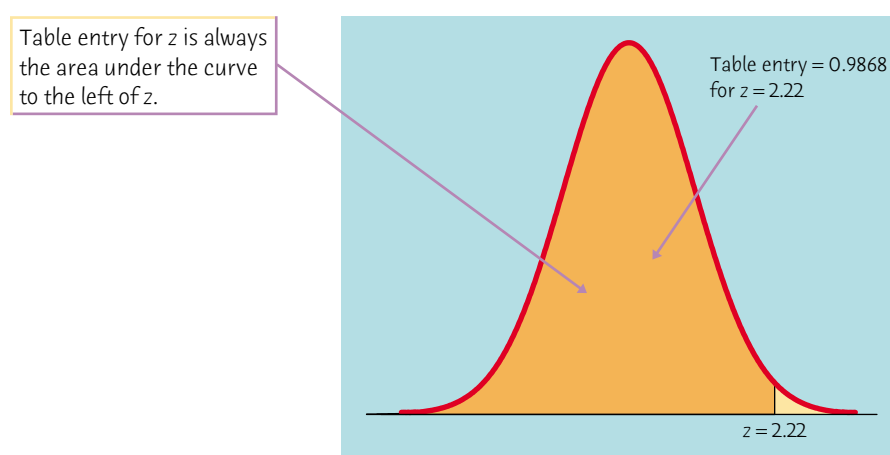
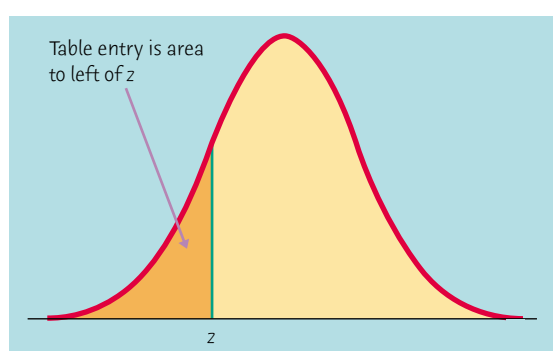


Figure 3.10(a) The area under a standard Normal curve to the left of the point $z = 2.22$ is 0.9868. Table A gives areas under the standard Normal curve.

THE STANDARD NORMAL TABLE

Table A is a table of areas under the standard Normal curve. The table entry for each value z is the area under the curve to the left of z .



EXAMPLE 3.5 Using the standard Normal table

Problem: Find the proportion of observations from the standard Normal distribution that are less than 2.22.

Solution: To find the area to the left of 2.22, locate 2.2 in the left-hand column of Table A, then locate the remaining digit 2 as .02 in the top row. The entry opposite 2.2 and under .02 is 0.9868. This is the area we seek. Figure 3.10(a) illustrates the relationship between the value $z = 2.22$ and the area 0.9868. Because $z = 2.22$ is the standardized value of height 70 inches, the proportion of young women who are less than 70 inches tall is 0.9868 (more than 98%).

Problem: Find the proportion of observations from the standard Normal distribution that are greater than -0.67 .

Solution: Enter Table A under $z = -0.67$. That is, find -0.6 in the left-hand column and .07 in the top row. The table entry is 0.2514. This is the area to the left of -0.67 . Because the total area under the curve is 1, the area lying to the right of -0.67 is $1 - 0.2514 = 0.7486$. Figure 3.10(b) illustrates these areas.

We can answer any question about proportions of observations in a Normal distribution by standardizing and then using the standard Normal table. Here is an outline of the method for finding the proportion of the distribution in any region.

FINDING NORMAL PROPORTIONS

1. State the problem in terms of the observed variable x .
2. Standardize x to restate the problem in terms of a standard Normal variable z . Draw a picture to show the area under the standard Normal curve.
3. Find the required area under the standard Normal curve, using Table A and the fact that the total area under the curve is 1.

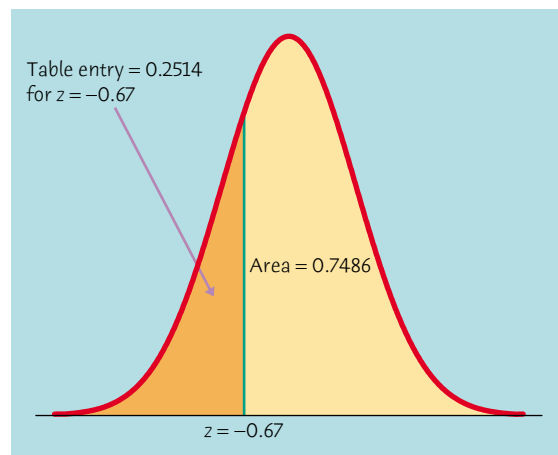


Figure 3.10(b) Areas under the standard Normal curve to the left and right of $z = -0.67$. Table A gives only areas to the left.

EXAMPLE 3.6 Normal distribution calculations

The level of cholesterol in the blood is important because high cholesterol levels may increase the risk of heart disease. The distribution of blood cholesterol levels in a large population of people of the same age and sex is roughly Normal. For 14-year-old boys, the mean is $\mu = 170$ milligrams of cholesterol per deciliter of blood (mg/dl) and the standard deviation is $\sigma = 30$ mg/dl.⁴ Levels above 240 mg/dl may require medical attention. What percent of 14-year-old boys have more than 240 mg/dl of cholesterol?

1. *State the problem.* Call the level of cholesterol in the blood x . The variable x has the $N(170, 30)$ distribution. We want the proportion of boys with $x > 240$.
2. *Standardize.* Subtract the mean, then divide by the standard deviation, to turn x into a standard Normal z :

$$\begin{aligned} x &> 240 \\ \frac{x - 170}{30} &> \frac{240 - 170}{30} \\ z &> 2.33 \end{aligned}$$

Figure 3.11 shows the standard Normal curve with the area of interest shaded.

3. *Use the table.* From Table A, we see that the proportion of observations less than 2.33 is 0.9901. About 99% of boys have cholesterol levels less than 240. The area to the right of 2.33 is therefore $1 - 0.9901 = 0.0099$. This is about 0.01, or 1%. Only about 1% of boys have high cholesterol.

In a Normal distribution, the proportion of observations with $x > 240$ is the same as the proportion with $x \geq 240$. There is no area under the curve and

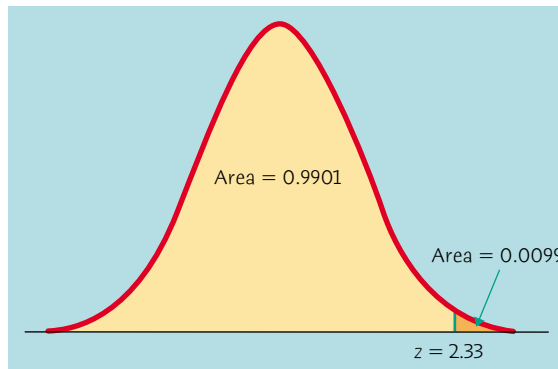


Figure 3.11 Areas under the standard Normal curve for Example 3.6.

exactly over 240, so the areas under the curve with $x > 240$ and $x \geq 240$ are the same. This isn't true of the actual data. There may be a boy with exactly 240 mg/dl of blood cholesterol. The Normal distribution is just an easy-to-use approximation, not a description of every detail in the actual data.

The key to using either software or Table A to do a Normal calculation is to sketch the area you want, then match that area with the areas that the table or software gives you. Here is another example.

EXAMPLE 3.7 More Normal distribution calculations

What percent of 14-year-old boys have blood cholesterol between 170 and 240 mg/dl?

1. *State the problem.* We want the proportion of boys with $170 \leq x \leq 240$.
2. *Standardize:*

$$\begin{aligned} 170 &\leq x \leq 240 \\ \frac{170 - 170}{30} &\leq \frac{x - 170}{30} \leq \frac{240 - 170}{30} \\ 0 &\leq z \leq 2.33 \end{aligned}$$

Figure 3.12 shows the area under the standard Normal curve.

3. *Use the table.* The area between 2.33 and 0 is the area below 2.33 *minus* the area below 0. Look at Figure 3.12 to check this. From Table A,

$$\begin{aligned} \text{area between 0 and 2.33} &= \text{area below 2.33} - \text{area below 0.00} \\ &= 0.9901 - 0.5000 = 0.4901 \end{aligned}$$

About 49% of boys have cholesterol levels between 170 and 240 mg/dl.

Sometimes we encounter a value of z more extreme than those appearing in Table A. For example, the area to the left of $z = -4$ is not given directly in the table. The z -values in Table A leave only area 0.0002 in each tail unaccounted

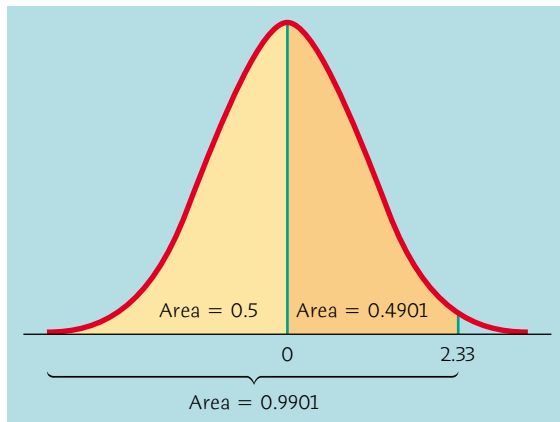


Figure 3.12 Areas under the standard Normal curve for Example 3.7.

for. For practical purposes, we can act as if there is zero area outside the range of Table A.

APPLY YOUR KNOWLEDGE

3.10 Use Table A to find the proportion of observations from a standard Normal distribution that satisfies each of the following statements. In each case, sketch a standard Normal curve and shade the area under the curve that is the answer to the question.

- (a) $z < 2.85$
- (b) $z > 2.85$
- (c) $z > -1.66$
- (d) $-1.66 < z < 2.85$

3.11 How hard do locomotives pull? An important measure of the performance of a locomotive is its “adhesion,” which is the locomotive’s pulling force as a multiple of its weight. The adhesion of one 4400-horsepower diesel locomotive model varies in actual use according to a Normal distribution with mean $\mu = 0.37$ and standard deviation $\sigma = 0.04$.

- (a) What proportion of adhesions measured in use are higher than 0.40?
- (b) What proportion of adhesions are between 0.40 and 0.50?
- (c) Improvements in the locomotive’s computer controls change the distribution of adhesion to a Normal distribution with mean $\mu = 0.41$ and standard deviation $\sigma = 0.02$. Find the proportions in (a) and (b) after this improvement.



(CORBIS)

Finding a value given a proportion

Examples 3.6 and 3.7 illustrate the use of Table A to find what proportion of the observations satisfies some condition, such as “blood cholesterol between

170 mg/dl and 240 mg/dl.” We may instead want to find the observed value with a given proportion of the observations above or below it. To do this, use Table A backward. Find the given proportion in the body of the table, read the corresponding z from the left column and top row, then “unstandardize” to get the observed value. Here is an example.

EXAMPLE 3.8 “Backward” Normal calculations

Scores on the SAT verbal test in 2002 followed approximately the $N(504, 111)$ distribution. How high must a student score in order to place in the top 10% of all students taking the SAT?

1. *State the problem.* We want to find the SAT score x with area 0.1 to its *right* under the Normal curve with mean $\mu = 504$ and standard deviation $\sigma = 111$. That’s the same as finding the SAT score x with area 0.9 to its *left*. Figure 3.13 poses the question in graphical form. Because Table A gives the areas to the left of z -values, always state the problem in terms of the area to the left of x .
2. *Use the table.* Look in the body of Table A for the entry closest to 0.9. It is 0.8997. This is the entry corresponding to $z = 1.28$. So $z = 1.28$ is the standardized value with area 0.9 to its left.
3. *Unstandardize* to transform the solution from the z back to the original x scale. We know that the standardized value of the unknown x is $z = 1.28$.

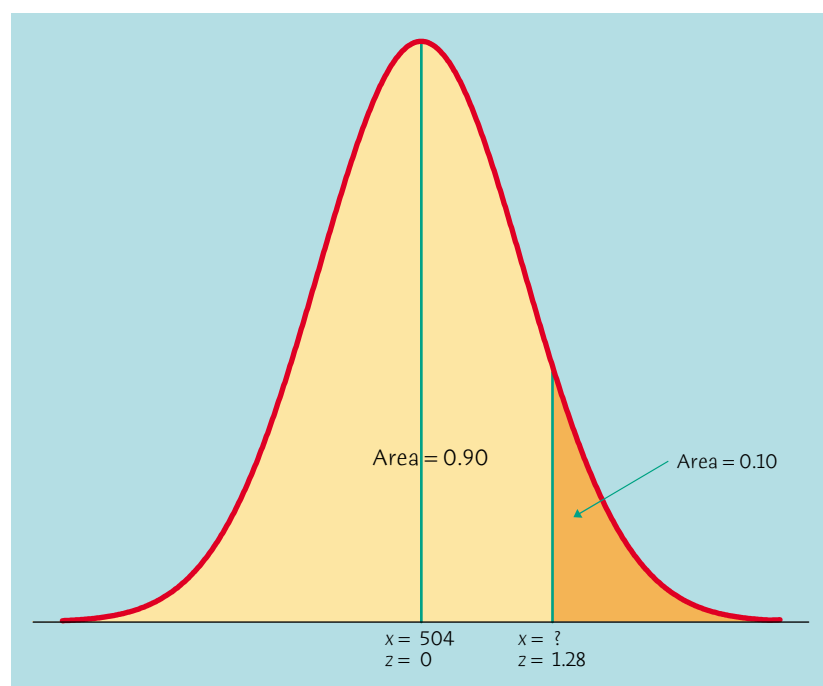


Figure 3.13 Locating the point on a Normal curve with area 0.10 to its right, for Example 3.8.

So x itself satisfies

$$\frac{x - 504}{111} = 1.28$$

Solving this equation for x gives

$$x = 504 + (1.28)(111) = 646.1$$

This equation should make sense: it says that x lies 1.28 standard deviations above the mean on this particular Normal curve. That is the “unstandardized” meaning of $z = 1.28$. We see that a student must score at least 647 to place in the highest 10%.

Here is the general formula for unstandardizing a z -score. To find the value x from the Normal distribution with mean μ and standard deviation σ corresponding to a given standard Normal value z , use

$$x = \mu + z\sigma$$

APPLY YOUR KNOWLEDGE

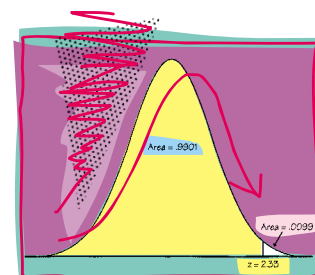
- 3.12** Use Table A to find the value z of a standard Normal variable that satisfies each of the following conditions. (Use the value of z from Table A that comes closest to satisfying the condition.) In each case, sketch a standard Normal curve with your value of z marked on the axis.
- (a) The point z with 25% of the observations falling below it.
 - (b) The point z with 40% of the observations falling above it.
- 3.13 IQ test scores.** Scores on the Wechsler Adult Intelligence Scale are approximately Normally distributed with $\mu = 100$ and $\sigma = 15$.
- (a) What IQ scores fall in the lowest 25% of the distribution?
 - (b) How high an IQ score is needed to be in the highest 5%?

Chapter 3 SUMMARY

We can sometimes describe the overall pattern of a distribution by a **density curve**. A density curve has total area 1 underneath it. An area under a density curve gives the proportion of observations that fall in a range of values.

A density curve is an idealized description of the overall pattern of a distribution that smooths out the irregularities in the actual data. We write the mean of a density curve as μ and the standard deviation of a density curve as σ to distinguish them from the mean $\bar{x}$ and standard deviation s of the actual data.

The mean, the median, and the quartiles of a density curve can be located by eye. The **mean** μ is the balance point of the curve. The **median** divides the area under the curve in half. The **quartiles** and the median divide the area



The bell curve?

Does the distribution of human intelligence follow the “bell curve” of a Normal distribution? Scores on IQ tests do roughly follow a Normal distribution. That is because a test score is calculated from a person’s answers in a way that is designed to produce a Normal distribution. To conclude that intelligence follows a bell curve, we must agree that the test scores directly measure intelligence. Many psychologists don’t think there is one human characteristic that we can call “intelligence” and can measure by a single test score.

under the curve into quarters. The **standard deviation** σ cannot be located by eye on most density curves.

The mean and median are equal for symmetric density curves. The mean of a skewed curve is located farther toward the long tail than is the median.

The **Normal distributions** are described by a special family of bell-shaped, symmetric density curves, called **Normal curves**. The mean μ and standard deviation σ completely specify a Normal distribution $N(\mu, \sigma)$. The mean is the center of the curve, and σ is the distance from μ to the change-of-curvature points on either side.

To **standardize** any observation x , subtract the mean of the distribution and then divide by the standard deviation. The resulting **z -score**

$$z = \frac{x - \mu}{\sigma}$$

says how many standard deviations x lies from the distribution mean.

All Normal distributions are the same when measurements are transformed to the standardized scale. In particular, all Normal distributions satisfy the **68–95–99.7 rule**, which describes what percent of observations lie within one, two, and three standard deviations of the mean.

If x has the $N(\mu, \sigma)$ distribution, then the **standardized variable** $z = (x - \mu)/\sigma$ has the **standard Normal distribution** $N(0, 1)$ with mean 0 and standard deviation 1. Table A gives the proportions of standard Normal observations that are less than z for many values of z . By standardizing, we can use Table A for any Normal distribution.

Chapter 3 EXERCISES

- 3.14** Figure 3.14 shows two Normal curves, both with mean 0. Approximately what is the standard deviation of each of these curves?

IQ test scores. The Wechsler Adult Intelligence Scale (WAIS) is the most common “IQ test.” The scale of scores is set separately for each age group and is approximately Normal with mean 100 and standard deviation 15. Use this distribution and the 68–95–99.7 rule to answer Exercises 3.15 to 3.17.



(Mel Yates/Taxi/Getty Images)

- 3.15** About what percent of people have WAIS scores
(a) above 100?
(b) above 145?
(c) below 85?
- 3.16 Retardation.** People with WAIS scores below 70 are considered mentally retarded when, for example, applying for Social Security disability benefits. What percent of adults are retarded by this criterion?
- 3.17 MENSA.** The organization MENSA, which calls itself “the high IQ society,” requires a WAIS score of 130 or higher for membership.

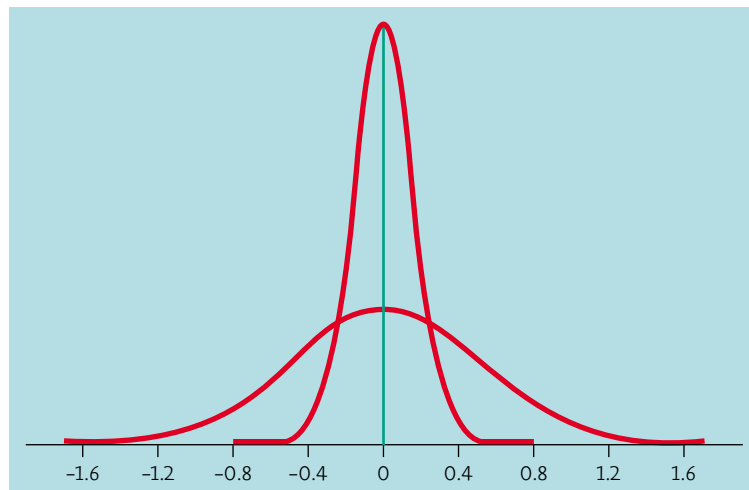


Figure 3.14 Two Normal curves with the same mean but different standard deviations, for Exercise 3.14.

(Similar scores on other tests are also accepted.) What percent of adults would qualify for membership?

3.18 Standard Normal drill. Use Table A to find the proportion of observations from a standard Normal distribution that falls in each of the following regions. In each case, sketch a standard Normal curve and shade the area representing the region.

- (a) $z \leq -2.25$
- (b) $z \geq -2.25$
- (c) $z > 1.77$
- (d) $-2.25 < z < 1.77$

3.19 Standard Normal drill.

- (a) Find the number z such that the proportion of observations that are less than z in a standard Normal distribution is 0.8.
- (b) Find the number z such that 35% of all observations from a standard Normal distribution are greater than z .

3.20 NCAA rules for athletes. The National Collegiate Athletic Association (NCAA) requires Division I athletes to score at least 820 on the combined mathematics and verbal parts of the SAT exam to compete in their first college year. (Higher scores are required for students with poor high school grades.) In 2002, the scores of the 1.3 million students taking the SATs were approximately Normal with mean 1020 and standard deviation 207. What percent of all students had scores less than 820?

3.21 More NCAA rules. The NCAA considers a student a “partial qualifier” eligible to practice and receive an athletic scholarship, but not

to compete, if the combined SAT score is at least 720. Use the information in the previous exercise to find the percent of all SAT scores that are less than 720.

3.22 Heights of men and women. The heights of women aged 20 to 29 follow approximately the $N(64, 2.7)$ distribution. Men the same age have heights distributed as $N(69.3, 2.8)$. What percent of young women are taller than the mean height of young men?

3.23 Heights of men and women. The heights of women aged 20 to 29 follow approximately the $N(64, 2.7)$ distribution. Men the same age have heights distributed as $N(69.3, 2.8)$. What percent of young men are shorter than the mean height of young women?



(Jim McGuire/Index Stock Imagery/PictureQuest)

3.24 Length of pregnancies. The length of human pregnancies from conception to birth varies according to a distribution that is approximately Normal with mean 266 days and standard deviation 16 days.

(a) What percent of pregnancies last less than 240 days (that's about 8 months)?

(b) What percent of pregnancies last between 240 and 270 days (roughly between 8 months and 9 months)?

(c) How long do the longest 20% of pregnancies last?

3.25 A surprising calculation. Changing the mean of a Normal distribution by a moderate amount can greatly change the percent of observations in the tails. Suppose that a college is looking for applicants with SAT math scores 750 and above.

(a) In 2002, the scores of men on the math SAT followed the $N(534, 116)$ distribution. What percent of men scored 750 or better?

(b) Women's SAT math scores that year had the $N(500, 110)$ distribution. What percent of women scored 750 or better? You see that the percent of men above 750 is almost three times the percent of women with such high scores.

3.26 Grading managers. Many companies “grade on a bell curve” to compare the performance of their managers and professional workers. This forces the use of some low performance ratings, so that not all workers are listed as “above average.” Ford Motor Company's “performance management process” for a time assigned 10% A grades, 80% B grades, and 10% C grades to the company's 18,000 managers. Suppose that Ford's performance scores really are Normally distributed. This year, managers with scores less than 25 received C's and those with scores above 475 received A's. What are the mean and standard deviation of the scores?

3.27 Weights aren't Normal. The heights of people of the same sex and similar ages follow Normal distributions reasonably closely. Weights, on

the other hand, are not Normally distributed. The weights of women aged 20 to 29 have mean 141.7 pounds and median 133.2 pounds. The first and third quartiles are 118.3 pounds and 157.3 pounds. What can you say about the shape of the weight distribution? Why?

3.28 Quartiles. The median of any Normal distribution is the same as its mean. We can use Normal calculations to find the quartiles for Normal distributions.

- (a) What is the area under the standard Normal curve to the left of the first quartile? Use this to find the value of the first quartile for a standard Normal distribution. Find the third quartile similarly.
- (b) Your work in (a) gives the z -scores for the quartiles of any Normal distribution. What are the quartiles for the lengths of human pregnancies? (Use the distribution in Exercise 3.24.)

3.29 Deciles. The *deciles* of any distribution are the points that mark off the lowest 10% and the highest 10%. On a density curve, these are the points with areas 0.1 and 0.9 to their left under the curve.

- (a) What are the deciles of the standard Normal distribution?
- (b) The heights of young women are approximately Normal with mean 64 inches and standard deviation 2.7 inches. What are the deciles of this distribution?

Chapter 3 MEDIA EXERCISES

The *Normal Curve* applet allows you to do Normal calculations quickly. It is somewhat limited by the number of pixels available for use, so that it can't hit every value exactly. In the exercises below, use the closest available values. In each case, *make a sketch* of the curve from the applet marked with the values you used to answer the questions asked.

3.30 How accurate is 68–95–99.7? The 68–95–99.7 rule for Normal distributions is a useful approximation. To see how accurate the rule is, drag one flag across the other so that the applet shows the area under the curve between the two flags.



- (a) Place the flags one standard deviation on either side of the mean. What is the area between these two values? What does the 68–95–99.7 rule say this area is?
- (b) Repeat for locations two and three standard deviations on either side of the mean. Again compare the 68–95–99.7 rule with the area given by the applet.

3.31 Where are the quartiles? How many standard deviations above and below the mean do the quartiles of any Normal distribution lie? (Use the standard Normal distribution to answer this question.)





3.32 Grading managers. In Exercise 3.26, we saw that Ford Motor Company grades its managers in such a way that the top 10% receive an A grade, the bottom 10% a C, and the middle 80% a B. Let's suppose that performance scores follow a Normal distribution. How many standard deviations above and below the mean do the A/B and B/C cutoffs lie? (Use the standard Normal distribution to answer this question.)



3.33 Do the data look Normal? The EESSEE stories provide data from many studies in various fields. Make a histogram of each of the following variables. In each case, do the data appear roughly Normal (at least single-peaked, symmetric, tails falling off on both side of the peak) or is their shape clearly not Normal? If the distribution is not Normal, briefly describe its shape.

- (a) The number of single owls per square kilometer, from "Habitat of the Spotted Owl."
- (b) Pretest Degree of Reading Power scores, from "Checkmating and Reading Skills."
- (c) Carbon monoxide emissions from the refinery (only), from "Emissions from an Oil Refinery."

CHAPTER

4



(Richard Hamilton Smith/Corbis)

Scatterplots and Correlation

A medical study finds that short women are more likely to have heart attacks than women of average height, while tall women have the fewest heart attacks. An insurance group reports that heavier cars have fewer deaths per 10,000 vehicles registered than do lighter cars. These and many other statistical studies look at the relationship between two variables. To understand such a relationship, we must often examine other variables as well. To conclude that shorter women have higher risk from heart attacks, for example, the researchers had to eliminate the effect of other variables such as weight and exercise habits. In this chapter we begin our study of relationships between variables. One of our main themes is that the relationship between two variables can be strongly influenced by other variables that are lurking in the background.

Explanatory and response variables

Because variation is everywhere, statistical relationships are overall tendencies, not ironclad rules. They allow individual exceptions. Although smokers on the average die younger than nonsmokers, some people live to 90 while smoking three packs a day. To study a relationship between two variables, we measure both variables on the same individuals. Often, we think that one of the variables explains or influences the other.

In this chapter we cover...

Explanatory and response variables

Displaying relationships: scatterplots

Interpreting scatterplots

Adding categorical variables to scatterplots

Measuring linear association: correlation

Facts about correlation

RESPONSE VARIABLE, EXPLANATORY VARIABLE

A **response variable** measures an outcome of a study. An **explanatory variable** explains or influences changes in a response variable.

independent variable
dependent variable

You will often find explanatory variables called **independent variables**, and response variables called **dependent variables**. The idea behind this language is that the response variable depends on the explanatory variable. Because “independent” and “dependent” have other meanings in statistics that are unrelated to the explanatory-response distinction, we prefer to avoid those words.

It is easiest to identify explanatory and response variables when we actually set values of one variable in order to see how it affects another variable.

EXAMPLE 4.1 The effects of alcohol

Alcohol has many effects on the body. One effect is a drop in body temperature. To study this effect, researchers give several different amounts of alcohol to mice, then measure the change in each mouse’s body temperature in the 15 minutes after taking the alcohol. Amount of alcohol is the explanatory variable, and change in body temperature is the response variable.

When we don’t set the values of either variable but just observe both variables, there may or may not be explanatory and response variables. Whether there are depends on how we plan to use the data.

EXAMPLE 4.2 Voting patterns

A political scientist looks at the Democrats’ share of the popular vote for president in each state in the two most recent presidential elections. She does not wish to explain one year’s data by the other’s, but rather to find a pattern that may shed light on political conditions. The political scientist has two related variables, but she does not use the explanatory-response distinction.

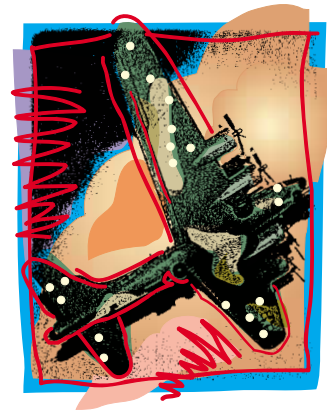
A political consultant looks at the same data with an eye to including the previous election result in a state as one of several explanatory variables in a model to predict how the state will vote in the next election. Now the earlier election result is an explanatory variable and the later result is a response variable.

In Example 4.1, alcohol actually *causes* a change in body temperature. There is no cause-and-effect relationship between votes in different elections in Example 4.2. If there is a close relationship, however, we can use a state’s percent Democratic in the last election to help predict the next election. Chapter 5 discusses prediction. Prediction requires that we identify an explanatory variable and a response variable. Some other statistical techniques ignore

this distinction. Do remember that calling one variable explanatory and the other response doesn't necessarily mean that changes in one *cause* changes in the other.

Most statistical studies examine data on more than one variable. Fortunately, statistical analysis of several-variable data builds on the tools we used to examine individual variables. The principles that guide our work also remain the same:

- First plot the data, then add numerical summaries.
- Look for overall patterns and deviations from those patterns.
- When the overall pattern is quite regular, use a compact mathematical model to describe it.



**After you plot
your data, think!**

The statistician Abraham Wald (1902–1950) worked on war problems during World War II. Wald invented some statistical methods that were military secrets until the war ended. Here is one of his simpler ideas. Asked where extra armor should be added to airplanes, Wald studied the location of enemy bullet holes in planes returning from combat. He plotted the locations on an outline of the plane. As data accumulated, most of the outline filled up. Put the armor in the few spots with no bullet holes, said Wald. That's where bullets hit the planes that didn't make it back.

APPLY YOUR KNOWLEDGE

- 4.1 Explanatory and response variables?** In each of the following situations, is it more reasonable to simply explore the relationship between the two variables or to view one of the variables as an explanatory variable and the other as a response variable? In the latter case, which is the explanatory variable and which is the response variable?
- (a) The amount of time a student spends studying for a statistics exam and the grade on the exam.
 - (b) The weight in kilograms and height in centimeters of a person.
 - (c) Inches of rain in the growing season and the yield of corn in bushels per acre.
 - (d) A student's scores on the SAT math exam and the SAT verbal exam.
 - (e) A family's income and the years of education their eldest child completes.
- 4.2 Coral reefs.** How sensitive to changes in water temperature are coral reefs? To find out, measure the growth of corals in aquariums where the water temperature is controlled at different levels. Growth is measured by weighing the coral before and after the experiment. What are the explanatory and response variables? Are they categorical or quantitative?
- 4.3 The risks of obesity.** A study observes a large group of people over a 10-year period. The goal is to see if overweight and obese people are more likely to die during the decade than people who weigh less. Such studies can be misleading, because obese people are more likely to be inactive and to be poor. What is the explanatory variable and the response variable? What other variables are mentioned that may influence the relationship between the explanatory variable and the response variable?

Displaying relationships: scatterplots

Relationships between two quantitative variables are best displayed graphically. The most useful graph for this purpose is a *scatterplot*.

EXAMPLE 4.3 State SAT scores

Some people use average SAT scores to rank state or local school systems. This is not proper, because the percent of high school students who take the SAT varies from place to place. Let's examine the relationship between the percent of a state's high school graduates who took the exam in 2002 and the state average SAT verbal score that year.

We think that “percent taking” will help explain “average score.” Therefore, “percent taking” is the explanatory variable and “average score” is the response variable. We want to see how average score changes when percent taking changes, so we put percent taking (the explanatory variable) on the horizontal axis. Figure 4.1 is the scatterplot. Each point represents a single state. In Colorado, for example, 28% took the SAT, and the average SAT verbal score was 543. Find 28 on the x (horizontal) axis and 543 on the y (vertical) axis. Colorado appears as the point (28, 543) above 28 and to the right of 543. Figure 4.1 shows how to locate Colorado's point on the plot.

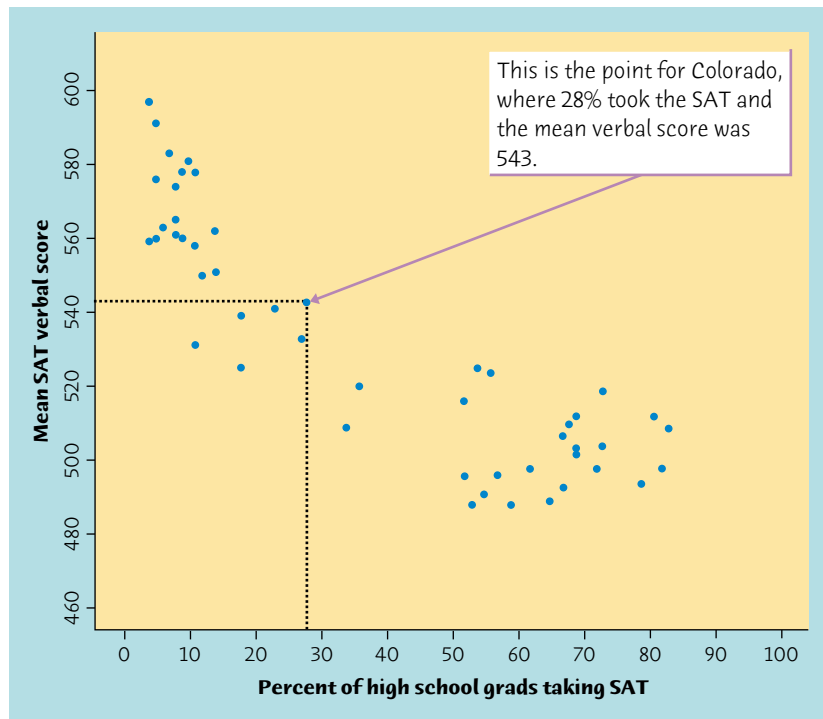


Figure 4.1 Scatterplot of the mean SAT verbal score in each state against the percent of that state's high school graduates who take the SAT. The dotted lines intersect at the point (28, 543), the data for Colorado.

SCATTERPLOT

A **scatterplot** shows the relationship between two quantitative variables measured on the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as the point in the plot fixed by the values of both variables for that individual.

Always plot the explanatory variable, if there is one, on the horizontal axis (the x axis) of a scatterplot. As a reminder, we usually call the explanatory variable x and the response variable y . If there is no explanatory-response distinction, either variable can go on the horizontal axis.

APPLY YOUR KNOWLEDGE

4.4 Bird colonies. One of nature’s patterns connects the percent of adult birds in a colony that return from the previous year and the number of new adults that join the colony. Here are data for 13 colonies of sparrowhawks:¹

Percent returning	New adults	Percent returning	New adults	Percent returning	New adults
74	5	62	15	46	18
66	6	52	16	60	19
81	8	45	17	46	20
52	11	62	18	38	20
73	12				



(Chris Beddall/Papilio/CORBIS)

Plot the count of new birds (response) against the percent of returning birds (explanatory).

Interpreting scatterplots

To interpret a scatterplot, apply the strategies of data analysis learned in Chapters 1 to 3.

EXAMINING A SCATTERPLOT

In any graph of data, look for the **overall pattern** and for striking **deviations** from that pattern.

You can describe the overall pattern of a scatterplot by the **form**, **direction**, and **strength** of the relationship.

An important kind of deviation is an **outlier**, an individual value that falls outside the overall pattern of the relationship.

clusters

EXAMPLE 4.4 Understanding state SAT scores

Figure 4.1 shows a clear *form*: there are two distinct **clusters** of states with a gap between them. In the cluster at the right of the plot, 52% or more of high school graduates take the SAT, and the average scores are low. The states in the cluster at the left have higher SAT scores and no more than 36% of graduates take the test. There are no clear outliers. That is, no points fall clearly outside the clusters.

What explains the clusters? There are two widely used college entrance exams, the SAT and the ACT Assessment exam. Each state favors one or the other. The left cluster in Figure 4.1 contains the ACT states, and the SAT states make up the right cluster. In ACT states, most students who take the SAT are applying to a selective college that requires SAT scores. This select group of students has a higher average score than the much larger group of students who take the SAT in SAT states.

The relationship in Figure 4.1 also has a clear *direction*: states in which a higher percent of students take the SAT tend to have lower average scores. This is a *negative association* between the two variables.

POSITIVE ASSOCIATION, NEGATIVE ASSOCIATION

Two variables are **positively associated** when above-average values of one tend to accompany above-average values of the other, and below-average values also tend to occur together.

Two variables are **negatively associated** when above-average values of one tend to accompany below-average values of the other, and vice versa.

The *strength* of a relationship in a scatterplot is determined by how closely the points follow a clear form. The overall relationship in Figure 4.1 is not strong—states with similar percents taking the SAT show quite a bit of scatter in their average scores. Here is an example of a stronger relationship with a clearer form.

EXAMPLE 4.5 Counting carnivores

Ecologists look at data to learn about nature’s patterns. One pattern they have found relates the size of a carnivore (body mass in kilograms) to how many of those carnivores there are in an area. The right measure of “how many” is to count carnivores per 10,000 kilograms of their prey in the area. Table 4.1 gives data for 25 carnivore species.²

To see the pattern, plot carnivore abundance (response) against body mass (explanatory). Biologists often find that patterns involving sizes and counts are simpler when we plot the logarithms of the data. Figure 4.2 does that—you can see that 1, 10, 100, and 1000 are equally spaced on the vertical scale.

This scatterplot shows a **moderately strong negative association**. Bigger carnivores are less abundant. The form of the association is **linear**. It is striking that animals from many different parts of the world should fit so simple a pattern. We could use the straight-line pattern to predict the abundance of another carnivore species from its body mass.

TABLE 4.1 Size and abundance of carnivores

Carnivore species	Body mass (kg)	Abundance	Carnivore species	Body mass (kg)	Abundance
Least weasel	0.14	1656.49	Eurasian lynx	20.0	0.46
Ermine	0.16	406.66	Wild dog	25.0	1.61
Small Indian mongoose	0.55	514.84	Dhole	25.0	0.81
Pine marten	1.3	31.84	Snow leopard	40.0	1.89
Kit fox	2.02	15.96	Wolf	46.0	0.62
Channel Island fox	2.16	145.94	Leopard	46.5	6.17
Arctic fox	3.19	21.63	Cheetah	50.0	2.29
Red fox	4.6	32.21	Puma	51.9	0.94
Bobcat	10.0	9.75	Spotted hyena	58.6	0.68
Canadian lynx	11.2	4.79	Lion	142.0	3.40
European badger	13.0	7.35	Tiger	181.0	0.33
Coyote	13.0	11.65	Polar bear	310.0	0.60
Ethiopian wolf	14.5	2.70			

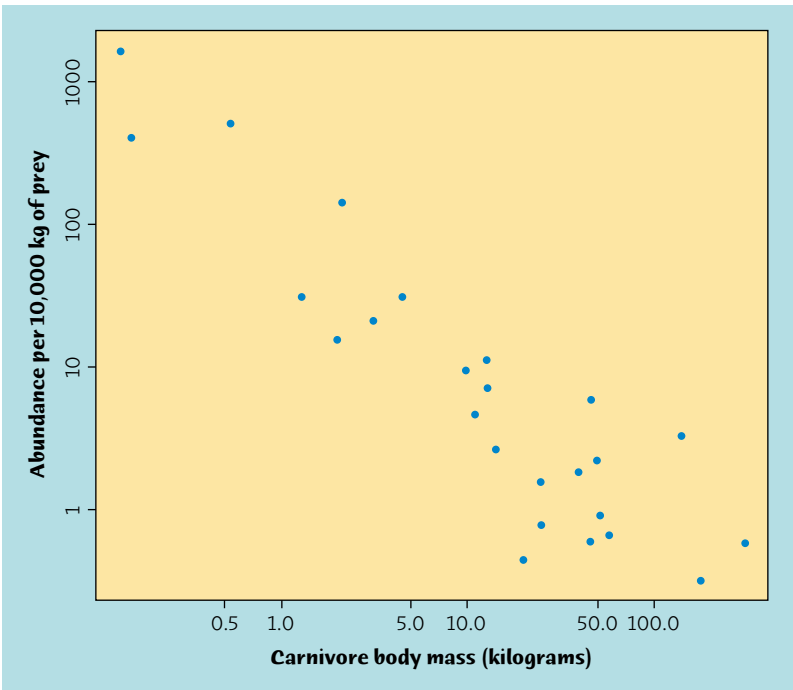


Figure 4.2 Scatterplot of the abundance of 25 species of carnivores against their body mass. Larger carnivores are less abundant. (Logarithmic scales are used for both variables.)

Of course, not all relationships have a simple form and a clear direction that we can describe as positive association or negative association. Exercise 4.6 gives an example that has no clear direction.

APPLY YOUR KNOWLEDGE

4.5 Bird colonies. Describe the form, direction, and strength of the relationship between number of new sparrowhawks in a colony and percent of returning adults, as displayed in your plot from Exercise 4.4.

For short-lived birds, the association between these variables is positive: changes in weather and food supply drive the populations of new and returning birds up or down together. For long-lived territorial birds, on the other hand, the association is negative because returning birds claim their territories in the colony and don't leave room for new recruits. Which type of species is the sparrowhawk?

4.6 Does fast driving waste fuel? How does the fuel consumption of a car change as its speed increases? Here are data for a British Ford Escort. Speed is measured in kilometers per hour, and fuel consumption is measured in liters of gasoline used per 100 kilometers traveled.³

Speed (km/h)	Fuel used (liters/100 km)	Speed (km/h)	Fuel used (liters/100 km)
10	21.00	90	7.57
20	13.00	100	8.27
30	10.00	110	9.03
40	8.00	120	9.87
50	7.00	130	10.79
60	5.90	140	11.77
70	6.30	150	12.83
80	6.95		

- Make a scatterplot. (Which is the explanatory variable?)
- Describe the form of the relationship. It is not linear. Explain why the form of the relationship makes sense.
- It does not make sense to describe the variables as either positively associated or negatively associated. Why?
- Is the relationship reasonably strong or quite weak? Explain your answer.

Adding categorical variables to scatterplots

The South long lagged behind the rest of the United States in the performance of its schools. Efforts to improve education have reduced the gap. We wonder if the South stands out in our study of state mean SAT scores.

EXAMPLE 4.6 The South has risen

Figure 4.3 enhances the scatterplot in Figure 4.1 by plotting 13 southern states with a different plot color. You see that the southern states blend in with the rest of

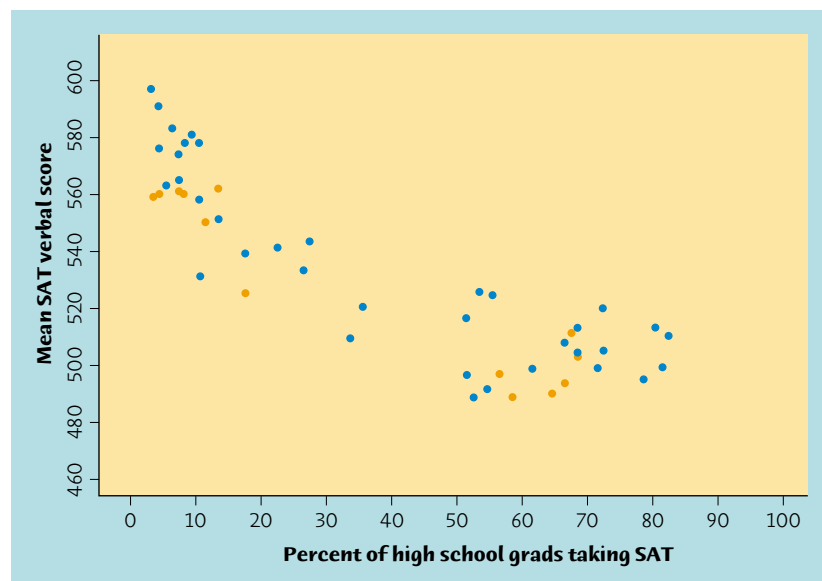


Figure 4.3 Mean SAT verbal score and percent of high school graduates who take the test, by state, with the southern states highlighted.

the country. Some are ACT states and some are SAT states, but neither group has markedly low SAT scores. The South no longer underperforms, at least as measured by SAT scores.

Dividing the states into “southern” and “nonsouthern” introduces a third variable into the scatterplot. This is a categorical variable that has only two values. The two values are displayed by the two different plotting colors.

CATEGORICAL VARIABLES IN SCATTERPLOTS

To add a categorical variable to a scatterplot, use a different plot color or symbol for each category.

APPLY YOUR KNOWLEDGE

- 4.7 How fast do icicles grow?** Japanese researchers measured the growth of icicles in a cold chamber under various conditions of temperature, wind, and water flow.⁴ Table 4.2 contains data produced under two sets of conditions. In both cases, there was no wind and the temperature was set at -11°C . Water flowed over the icicle at a higher rate (29.6 milligrams per second) in run 8905 and at a slower rate (11.9 mg/s) in run 8903.
- Make a scatterplot of the length of the icicle in centimeters versus time in minutes, using separate symbols or colors for the two runs.
 - What does your plot show about the pattern of growth of icicles? What does it show about the effect of changing the rate of water flow on icicle growth?

TABLE 4.2 Growth of icicles over time

Run 8903				Run 8905			
Time (min)	Length (cm)	Time (min)	Length (cm)	Time (min)	Length (cm)	Time (min)	Length (cm)
10	0.6	130	18.1	10	0.3	130	10.4
20	1.8	140	19.9	20	0.6	140	11.0
30	2.9	150	21.0	30	1.0	150	11.9
40	4.0	160	23.4	40	1.3	160	12.7
50	5.0	170	24.7	50	3.2	170	13.9
60	6.1	180	27.8	60	4.0	180	14.6
70	7.9			70	5.3	190	15.8
80	10.1			80	6.0	200	16.2
90	10.9			90	6.9	210	17.9
100	12.7			100	7.8	220	18.8
110	14.4			110	8.3	230	19.9
120	16.6			120	9.6	240	21.1

Measuring linear association: correlation

A scatterplot displays the form, direction, and strength of the relationship between two quantitative variables. Linear relations are particularly important because a straight line is a simple pattern that is quite common. We say a linear relation is strong if the points lie close to a straight line, and weak if they are widely scattered about a line. Our eyes are not good judges of how strong a linear relationship is. The two scatterplots in Figure 4.4 depict exactly the same data, but the lower plot is drawn smaller in a large field. The lower plot seems to show a stronger linear relationship. Our eyes can be fooled by changing the plotting scales or the amount of white space around the cloud of points in a scatterplot.⁵ We need to follow our strategy for data analysis by using a numerical measure to supplement the graph. *Correlation* is the measure we use.

CORRELATION

The **correlation** measures the direction and strength of the linear relationship between two quantitative variables. Correlation is usually written as r .

Suppose that we have data on variables x and y for n individuals. The values for the first individual are x_1 and y_1 , the values for the second individual are x_2 and y_2 , and so on. The means and standard deviations of the two variables are $\bar{x}$ and s_x for the x -values, and $\bar{y}$ and s_y for the y -values. The correlation r between x and y is

$$r = \frac{1}{n - 1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

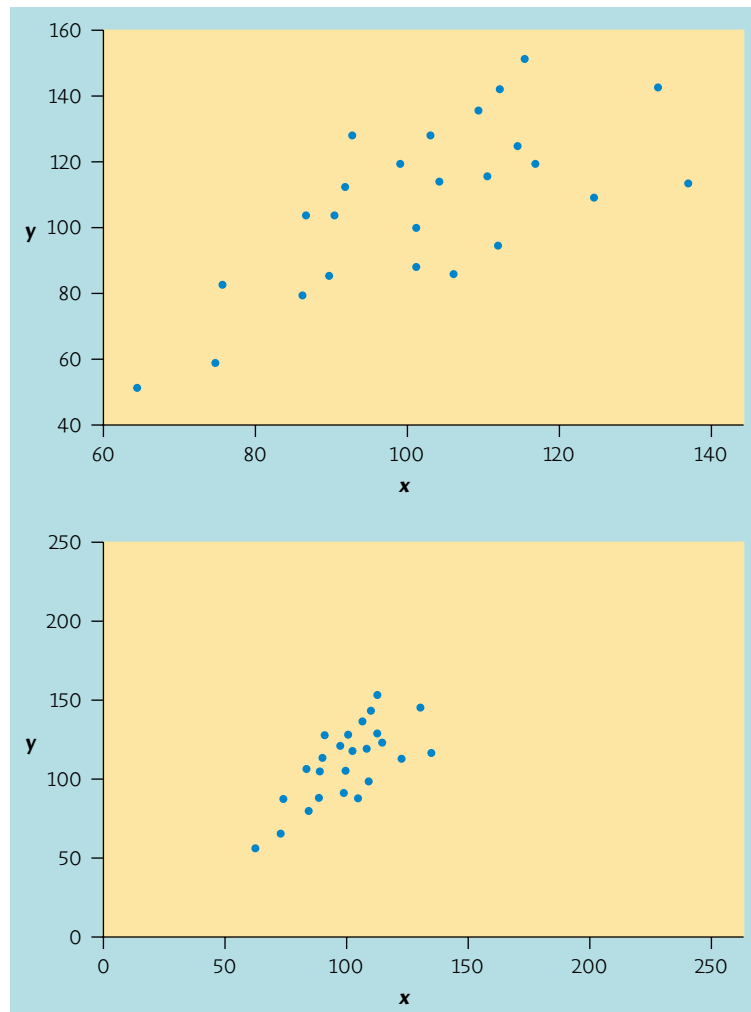


Figure 4.4 Two scatterplots of the same data. The straight-line pattern in the lower plot appears stronger because of the surrounding white space.

As always, the summation sign Σ means “add these terms for all the individuals.” The formula for the correlation r is a bit complex. It helps us see what correlation is, but in practice you should use software or a calculator that finds r from keyed-in values of two variables x and y . Exercise 4.8 asks you to calculate a correlation step-by-step from the definition to solidify its meaning.

The formula for r begins by standardizing the observations. Suppose, for example, that x is height in centimeters and y is weight in kilograms and that we have height and weight measurements for n people. Then $\bar{x}$ and s_x are the mean and standard deviation of the n heights, both in centimeters. The value

$$\frac{x_i - \bar{x}}{s_x}$$



Death from superstition?

Is there a relationship between superstitious beliefs and bad things happening? Apparently there is. Chinese and Japanese people think the number 4 is unlucky, because when pronounced it sounds like the word for “death.” Sociologists looked at 15 years’ of death certificates for Chinese and Japanese Americans and for white Americans. Deaths from heart disease were notably higher on the fourth day of the month among Chinese and Japanese, but not among whites. The sociologists think the explanation is increased stress on “unlucky days.”

is the standardized height of the i th person, familiar from Chapter 1. The standardized height says how many standard deviations above or below the mean a person's height lies. Standardized values have no units—in this example, they are no longer measured in centimeters. Standardize the weights also. The correlation r is an average of the products of the standardized height and the standardized weight for the n people.

APPLY YOUR KNOWLEDGE

4.8 Classifying fossils. *Archaeopteryx* is an extinct beast having feathers like a bird but teeth and a long bony tail like a reptile. Only six fossil specimens are known. Because these specimens differ greatly in size, some scientists think they belong to different species. We will examine some data. If the specimens belong to the same species and differ in size because some are younger than others, there should be a positive linear relationship between the lengths of a pair of bones from all individuals. An outlier from this relationship would suggest a different species. Here are data on the lengths in centimeters of the femur (a leg bone) and the humerus (a bone in the upper arm) for the five specimens that preserve both bones:⁶

Femur	38	56	59	64	74
Humerus	41	63	70	72	84

- (a) Make a scatterplot. Do you think that all five specimens come from the same species?
- (b) Find the correlation r step-by-step. That is, find the mean and standard deviation of the femur lengths and of the humerus lengths. (Use your calculator for means and standard deviations.) Then find the five standardized values for each variable and use the formula for r .
- (c) Now enter these data into your calculator and use the calculator's correlation function to find r . Check that you get the same result as in (b).

Facts about correlation

The formula for correlation helps us see that r is positive when there is a positive association between the variables. Height and weight, for example, have a positive association. People who are above average in height tend to also be above average in weight. Both the standardized height and the standardized weight are positive. People who are below average in height tend to also have

below-average weight. Then both standardized height and standardized weight are negative. In both cases, the products in the formula for r are mostly positive and so r is positive. In the same way, we can see that r is negative when the association between x and y is negative. More detailed study of the formula gives more detailed properties of r . Here is what you need to know in order to interpret correlation.

1. Correlation makes no distinction between explanatory and response variables. It makes no difference which variable you call x and which you call y in calculating the correlation.
2. Correlation requires that both variables be quantitative, so that it makes sense to do the arithmetic indicated by the formula for r . We cannot calculate a correlation between the incomes of a group of people and what city they live in, because city is a categorical variable.
3. Because r uses the standardized values of the observations, r does not change when we change the units of measurement of x , y , or both. Measuring height in inches rather than centimeters and weight in pounds rather than kilograms does not change the correlation between height and weight. The correlation r itself has no unit of measurement; it is just a number.
4. Positive r indicates positive association between the variables, and negative r indicates negative association.
5. The correlation r is always a number between -1 and 1 . Values of r near 0 indicate a very weak linear relationship. The strength of the linear relationship increases as r moves away from 0 toward either -1 or 1 . Values of r close to -1 or 1 indicate that the points in a scatterplot lie close to a straight line. The extreme values $r = -1$ and $r = 1$ occur only in the case of a perfect linear relationship, when the points lie exactly along a straight line.
6. Correlation measures the strength of only a linear relationship between two variables. Correlation does not describe curved relationships between variables, no matter how strong they are.
7. Like the mean and standard deviation, the correlation is not resistant: r is strongly affected by a few outlying observations. Use r with caution when outliers appear in the scatterplot.

The scatterplots in Figure 4.5 illustrate how values of r closer to 1 or -1 correspond to stronger linear relationships. To make the meaning of r clearer, the standard deviations of both variables in these plots are equal, and the horizontal and vertical scales are the same. In general, it is not so easy to guess the value of r from the appearance of a scatterplot. Remember that changing the plotting scales in a scatterplot may mislead our eyes, but it does not change the correlation. To explore how extreme observations can influence r , use the *Correlation and Regression* applet.





Figure 4.5 How correlation measures the strength of a linear relationship. Patterns closer to a straight line have correlations closer to 1 or -1 .

The real data we have examined also illustrate how correlation measures the strength and direction of linear relationships. Figure 4.2 shows a strong negative linear relationship between the logarithms of body mass and abundance for carnivore species. The correlation is $r = -0.9124$. Figure 4.1 shows a weaker but still quite strong negative association between percent of students taking the SAT and the mean SAT verbal score in a state. The correlation is $r = -0.8814$.

Do remember that **correlation is not a complete description of two-variable data**, even when the relationship between the variables is linear. You should give the means and standard deviations of both x and y along with the correlation. (Because the formula for correlation uses the means and standard deviations, these measures are the proper choice to accompany a correlation.) Conclusions based on correlations alone may require rethinking in the light of a more complete description of the data.

EXAMPLE 4.7 Scoring figure skaters

Each phase of a figure-skating competition is scored by a panel of judges who use a scale from 1 to 6. The subjective nature of the scoring often results in controversy. We have the scores awarded by two judges, Marina and Steve, for a large number of performances. How well do they agree? Calculation shows that the correlation between their scores is $r = 0.9$, but the mean of Marina's scores is 0.5 point lower than Steve's mean.

These facts do not contradict each other. They are simply different kinds of information. The mean scores show that Marina awards lower scores than Steve. But because Marina gives *every* skater a score about 0.5 point lower than Steve, the correlation remains high. Adding or subtracting the same number to all values of either x or y does not change the correlation. If Marina and Steve both rate all skaters, the contest is fairly scored because they agree on which performances are better than others. The high r shows their agreement. But if Marina scores one skater and Steve another, we must add 0.5 point to Marina's score to arrive at a fair comparison.

APPLY YOUR KNOWLEDGE

- 4.9 Changing the correlation.
- (a) Use your calculator to find the correlation between the percent of returning birds and the number of new birds from the data in Exercise 4.4.
 - (b) Make a scatterplot of the data with two new points added. Point A: 10% return, 25 new birds. Point B: 40% return, 5 new birds. Find two new correlations, for the original data plus Point A, and for the original data plus Point B.
 - (c) Explain in terms of what correlation measures why adding Point A makes the correlation stronger (closer to -1) and adding Point B makes the correlation weaker (closer to 0).
- 4.10 Wives and husbands. If women always married men who were 2 years older than themselves, what would be the correlation between the ages of husband and wife? (Hint: Draw a scatterplot for several ages.)
- 4.11 Strong association but no correlation. The gas mileage of an automobile first increases and then decreases as the speed increases. Suppose that this relationship is very regular, as shown by the following data on speed (miles per hour) and mileage (miles per gallon):

Speed	20	30	40	50	60
MPG	24	28	30	28	24

Make a scatterplot of mileage versus speed. Show that the correlation between speed and mileage is $r = 0$. Explain why the correlation is 0 even though there is a strong relationship between speed and mileage.

Chapter 4 SUMMARY

To study relationships between variables, we must measure the variables on the same group of individuals.

If we think that a variable x may explain or even cause changes in another variable y , we call x an **explanatory variable** and y a **response variable**.

A **scatterplot** displays the relationship between two quantitative variables measured on the same individuals. Mark values of one variable on the horizontal axis (x axis) and values of the other variable on the vertical axis (y axis). Plot each individual's data as a point on the graph. Always plot the explanatory variable, if there is one, on the x axis of a scatterplot.

Plot points with different colors or symbols to see the effect of a categorical variable in a scatterplot.

In examining a scatterplot, look for an overall pattern showing the **form**, **direction**, and **strength** of the relationship, and then for **outliers** or other deviations from this pattern.

Form: **Linear relationships**, where the points show a straight-line pattern, are an important form of relationship between two variables. Curved relationships and **clusters** are other forms to watch for.

Direction: If the relationship has a clear direction, we speak of either **positive association** (high values of the two variables tend to occur together) or **negative association** (high values of one variable tend to occur with low values of the other variable.)

Strength: The **strength** of a relationship is determined by how close the points in the scatterplot lie to a simple form such as a line.

The **correlation r** measures the strength and direction of the linear association between two quantitative variables x and y . Although you can calculate a correlation for any scatterplot, r measures only straight-line relationships.

Correlation indicates the direction of a linear relationship by its sign: $r > 0$ for a positive association and $r < 0$ for a negative association. Correlation always satisfies $-1 \leq r \leq 1$ and indicates the strength of a relationship by how close it is to -1 or 1 . Perfect correlation, $r = \pm 1$, occurs only when the points on a scatterplot lie exactly on a straight line.

Correlation ignores the distinction between explanatory and response variables. The value of r is not affected by changes in the unit of measurement of either variable. Correlation is not resistant, so outliers can greatly change the value of r .

Chapter 4 EXERCISES

4.12 Sports car gas mileage. Table 1.2 (page 12) gives the city and highway gas mileages for two-seater cars. We expect a positive association

between the city and highway mileages of a group of vehicles. Make a scatterplot that shows the relationship between city and highway mileage, using city mileage as the explanatory variable. Describe the overall pattern. Does the outlier (the Honda Insight) extend the pattern of the other cars or is it far from the line they form?

4.13 IQ and school grades. Do students with higher IQ test scores tend to do better in school? Figure 4.6 is a scatterplot of IQ and school grade point average (GPA) for all 78 seventh-grade students in a rural Midwest school.⁷

- Say in words what a positive association between IQ and GPA would mean. Does the plot show a positive association?
- What is the form of the relationship? Is it roughly linear? Is it very strong? Explain your answers.
- At the bottom of the plot are several points that we might call outliers. One student in particular has a very low GPA despite an average IQ score. What are the approximate IQ and GPA for this student?

4.14 Sports car gas mileage. Find the correlation between city and highway mileage in Table 1.2, leaving out the Insight. Explain how the correlation matches the pattern of the plot. Based on your plot in Exercise 4.12, will adding the Insight make the correlation stronger

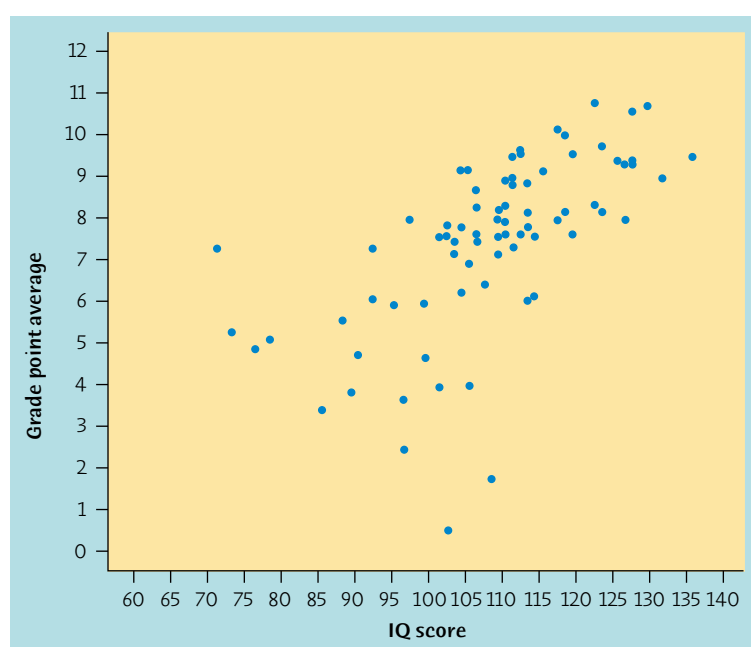


Figure 4.6 Scatterplot of school grade point average versus IQ test score for seventh-grade students, for Exercise 4.13.

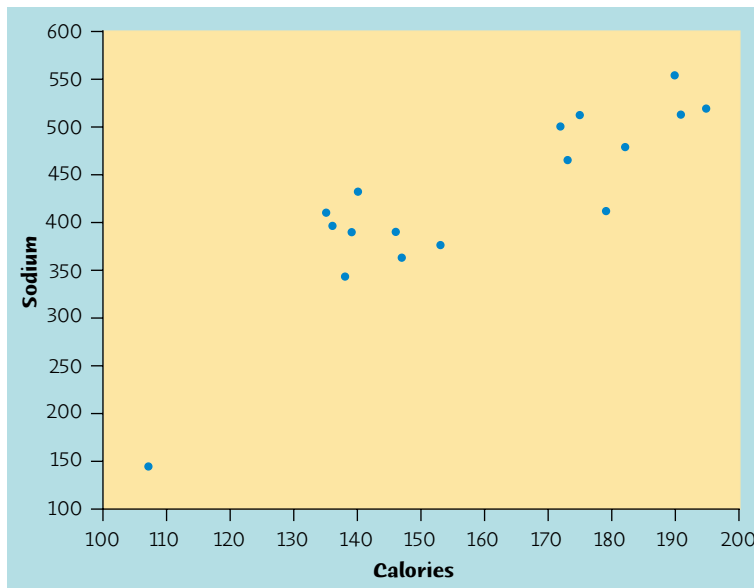


Figure 4.7 Scatterplot of milligrams of sodium and calories in each of 17 brands of meat hot dogs, for Exercise 4.15.

(closer to 1) or weaker? Verify your guess by calculating the correlation for all 22 cars.

4.15 Calories and salt in hot dogs. Are hot dogs that are high in calories also high in salt? Figure 4.7 is a scatterplot of the calories and salt content (measured as milligrams of sodium) in 17 brands of meat hot dogs.⁸

- Roughly what are the lowest and highest calorie counts among these brands? Roughly what is the sodium level in the brands with the fewest and with the most calories?
- Does the scatterplot show a clear positive or negative association? Say in words what this association means about calories and salt in hot dogs.
- Are there any outliers? Is the relationship (ignoring any outliers) roughly linear in form? Still ignoring any outliers, how strong would you say the relationship between calories and sodium is?

4.16 Heating a home. The Sanchez household is about to install solar panels to reduce the cost of heating their house. In order to know how much the solar panels help, they record their consumption of natural gas before the panels are installed. Gas consumption is higher in cold weather, so the relationship between outside temperature and gas consumption is important. Here are data for 16 consecutive months:⁹

Month	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	June
Degree-days	24	51	43	33	26	13	4	0
Gas used	6.3	10.9	8.9	7.5	5.3	4.0	1.7	1.2
Month	July	Aug.	Sept.	Oct.	Nov.	Dec.	Jan.	Feb.
Degree-days	0	1	6	12	30	32	52	30
Gas used	1.2	1.2	2.1	3.1	6.4	7.2	11.0	6.9

Outside temperature is recorded in degree-days, a common measure of demand for heating. A day's degree-days are the number of degrees its average temperature falls below 65° F. Gas used is recorded in hundreds of cubic feet. Make a plot and describe the pattern. Is correlation a reasonable way to describe the pattern? What is the correlation?

- 4.17 Thinking about correlation.** Figure 4.6 is a scatterplot of school grade point average versus IQ score for 78 seventh-grade students.
- (a) Is the correlation r for these data near -1 , clearly negative but not near -1 , near 0, clearly positive but not near 1, or near 1? Explain your answer.
- (b) Figure 4.7 shows the calories and sodium content in 17 brands of meat hot dogs. Is the correlation here closer to 1 than that for Figure 4.6, or closer to zero? Explain your answer.

- 4.18 Do solar panels reduce gas usage?** After the Sanchez household gathered the information recorded in Exercise 4.16, they added solar panels to their house. They then measured their natural-gas consumption for 23 more months. Here are the data:¹⁰

Degree-days	19	3	3	0	0	0	8	11	27	46	38	34
Gas	3.2	2.0	1.6	1.0	0.7	0.7	1.6	3.1	5.1	7.7	7.0	6.1
Degree-days	16	9	2	1	0	2	3	18	32	34	40	
Gas	3.0	2.1	1.3	1.0	1.0	1.0	1.2	3.4	6.1	6.5	7.5	

Add the new data to your scatterplot from Exercise 4.16, using a different color or symbol. What do the before-and-after data show about the effect of solar panels?

- 4.19 Thinking about correlation.** Both Figures 4.6 and 4.7 contain outliers. Removing the outliers will *increase* the correlation r in one figure and *decrease* r in the other figure. What happens in each figure, and why?
- 4.20 Merlins breeding.** Often the percent of an animal species in the wild that survive to breed again is lower following a successful breeding season. This is part of nature's self-regulation to keep population size stable. A study of merlins (small falcons) in northern Sweden observed



(Russell Burden/Index Stock Imagery/PictureQuest)

the number of breeding pairs in an isolated area and the percent of males (banded for identification) who returned the next breeding season. Here are data for nine years:¹¹

Breeding pairs	Percent of males returning
28	82
29	83, 70, 61
30	69
32	58
33	43
38	50, 47

- (a) Why is the response variable the *percent* of males that return rather than the *number* of males that return?
- (b) Make a scatterplot. Describe the pattern. Do the data support the theory that a smaller percent of birds survive following a successful breeding season?



(Julian Calder/Stone/Getty Images)

4.21 Mutual-fund performance. Many mutual funds compare their performance with that of a benchmark, an index of the returns on all securities of the kind the fund buys. The Vanguard International Growth Fund, for example, takes as its benchmark the Morgan Stanley EAFE (Europe, Australasia, Far East) index of overseas stock market performance. Here are the percent returns for the fund and for the EAFE from 1982 (the first full year of the fund’s existence) to 2001:¹²

Year	Fund	EAFE	Year	Fund	EAFE
1982	5.27	−1.86	1992	−5.79	−12.17
1983	43.08	23.69	1993	44.74	32.56
1984	−1.02	7.38	1994	0.76	7.78
1985	56.94	56.16	1995	14.89	11.21
1986	56.71	69.44	1996	14.65	6.05
1987	12.48	24.63	1997	4.12	1.78
1988	11.61	28.27	1998	16.93	20.00
1989	24.76	10.54	1999	26.34	26.96
1990	−12.05	−23.45	2000	−8.60	−14.17
1991	4.74	12.13	2001	−18.92	−21.44

Make a scatterplot suitable for predicting fund returns from EAFE returns. Is there a clear straight-line pattern? How strong is this pattern? (Give a numerical measure.) Are there any extreme outliers from the straight-line pattern?

- 4.22 Mice.** For a biology project, you measure the tail length (centimeters) and weight (grams) of 12 mice of the same variety.
- Explain why you expect the correlation between tail length and weight to be positive.
 - The mean tail length turns out to be 9.8 centimeters. What is the mean length in inches? (There are 2.54 centimeters in an inch.)
 - The correlation between tail length and weight turns out to be $r = 0.6$. If you measured length in inches instead of centimeters, what would be the new value of r ?
- 4.23 Statistics for investing.** Investment reports now often include correlations. Following a table of correlations among mutual funds, a report adds: “Two funds can have perfect correlation, yet different levels of risk. For example, Fund A and Fund B may be perfectly correlated, yet Fund A moves 20% whenever Fund B moves 10%.” Write a brief explanation, for someone who knows no statistics, of how this can happen. Include a sketch to illustrate your explanation.
- 4.24 Nematodes and tomatoes.** To demonstrate the effect of nematodes (microscopic worms) on plant growth, a student introduces different numbers of nematodes into 16 planting pots. He then transplants a tomato seedling into each pot. Here are data on the increase in height of the seedlings (in centimeters) 14 days after planting:¹³

Nematodes	Seedling growth			
0	10.8	9.1	13.5	9.2
1,000	11.1	11.1	8.2	11.3
5,000	5.4	4.6	7.4	5.0
10,000	5.8	5.3	3.2	7.5

- Make a scatterplot of the response variable (growth) against the explanatory variable (nematode count). Then compute the mean growth for each group of seedlings, plot the means against the nematode counts, and connect these four points with line segments.
 - Briefly describe the conclusions about the effects of nematodes on plant growth that these data suggest.
- 4.25 How many corn plants are too many?** How much corn per acre should a farmer plant to obtain the highest yield? Too few plants will give a low yield. On the other hand, if there are too many plants, they will compete with each other for moisture and nutrients, and yields will fall. To find the best planting rate, plant at different rates on several plots of



(Alvis Upitis/The Image Bank/Getty Images)

ground and measure the harvest. (Be sure to treat all the plots the same except for the planting rate.) Here are data from such an experiment:¹⁴

Plants per acre	Yield (bushels per acre)			
12,000	150.1	113.0	118.4	142.6
16,000	166.9	120.7	135.2	149.8
20,000	165.3	130.1	139.6	149.9
24,000	134.7	138.4	156.1	
28,000	119.0	150.5		

- (a) Is yield or planting rate the explanatory variable?
- (b) Make a scatterplot of yield and planting rate. Use a scale of yields from 100 to 200 bushels per acre so that the pattern will be clear.
- (c) Describe the overall pattern of the relationship. Is it linear? Is there a positive or negative association, or neither?
- (d) Find the mean yield for each of the five planting rates. Plot each mean yield against its planting rate on your scatterplot and connect these five points with lines. This combination of numerical description and graphing makes the relationship clearer. What planting rate would you recommend to a farmer whose conditions were similar to those in the experiment?

4.26 The effect of changing units. Changing the units of measurement can dramatically alter the appearance of a scatterplot. Return to the fossil data from Exercise 4.8:

Femur	38	56	59	64	74
Humerus	41	63	70	72	84

These measurements are in centimeters. Suppose a mad scientist measured the femur in meters and the humerus in millimeters. The data would then be

Femur	0.38	0.56	0.59	0.64	0.74
Humerus	410	630	700	720	840

- (a) Draw an x axis from 0 to 75 and a y axis from 0 to 850. Plot the original data on these axes. Then plot the new data on the same axes in a different color. The two plots look very different.
- (b) Nonetheless, the correlation is exactly the same for the two sets of measurements. Why do you know that this is true without doing any calculations? Find the two correlations to verify that they are the same.

TABLE 4.3 Wine consumption and heart attacks

Country	Alcohol from wine	Heart disease deaths	Country	Alcohol from wine	Heart disease deaths
Australia	2.5	211	Netherlands	1.8	167
Austria	3.9	167	New Zealand	1.9	266
Belgium	2.9	131	Norway	0.8	227
Canada	2.4	191	Spain	6.5	86
Denmark	2.9	220	Sweden	1.6	207
Finland	0.8	297	Switzerland	5.8	115
France	9.1	71	United Kingdom	1.3	285
Iceland	0.8	211	United States	1.2	199
Ireland	0.7	300	West Germany	2.7	172
Italy	7.9	107			

- 4.27 Is wine good for your heart?** There is some evidence that drinking moderate amounts of wine helps prevent heart attacks. Table 4.3 gives data on yearly wine consumption (liters of alcohol from drinking wine, per person) and yearly deaths from heart disease (deaths per 100,000 people) in 19 developed nations.¹⁵
- (a) Make a scatterplot that shows how national wine consumption helps explain heart disease death rates.
 - (b) Describe the form of the relationship. Is there a linear pattern? How strong is the relationship?
 - (c) Is the direction of the association positive or negative? Explain in simple language what this says about wine and heart disease. Do you think these data give good evidence that drinking wine *causes* a reduction in heart disease deaths? Why?
- 4.28 Teaching and research.** A college newspaper interviews a psychologist about student ratings of the teaching of faculty members. The psychologist says, “The evidence indicates that the correlation between the research productivity and teaching rating of faculty members is close to zero.” The paper reports this as “Professor McDaniel said that good researchers tend to be poor teachers, and vice versa.” Explain why the paper’s report is wrong. Write a statement in plain language (don’t use the word “correlation”) to explain the psychologist’s meaning.
- 4.29 Investment diversification.** A mutual-fund company’s newsletter says, “A well-diversified portfolio includes assets with low correlations.” The newsletter includes a table of correlations between the returns on various classes of investments. For example, the correlation between municipal bonds and large-cap stocks is 0.50, and the correlation between municipal bonds and small-cap stocks is 0.21.

102 CHAPTER 4 • Scatterplots and Correlation

- (a) Rachel invests heavily in municipal bonds. She wants to diversify by adding an investment whose returns do not closely follow the returns on her bonds. Should she choose large-cap stocks or small-cap stocks for this purpose? Explain your answer.
- (b) If Rachel wants an investment that tends to increase when the return on her bonds drops, what kind of correlation should she look for?

4.30 Sloppy writing about correlation. Each of the following statements contains a blunder. Explain in each case what is wrong.

- (a) “There is a high correlation between the gender of American workers and their income.”
- (b) “We found a high correlation ($r = 1.09$) between students’ ratings of faculty teaching and ratings made by other faculty members.”
- (c) “The correlation between planting rate and yield of corn was found to be $r = 0.23$ bushel.”

Chapter 4 MEDIA EXERCISES



4.31 SAT scores. The data plotted in Figure 4.1 came from the Web site of the College Board, www.collegeboard.com. At first glance, this site seems to be of interest mainly to students applying to college and planning to take the SAT. If you can locate it, there is a mine of information about SAT scores. Try clicking on “For Education Professionals” and then on “College Bound Seniors.” You can then access a list of data tables. Use data from these to do *one* of these tasks:

- (a) Make a time plot of the average SAT verbal score of all students for as many past years as the data table allows. What trend or other pattern do you see? Has the trend changed over time?
- (b) Make a scatterplot of the mean SAT verbal and math scores for the states in a recent year. What overall pattern do you see? What is the correlation? Hawaii is often a mild outlier. What might explain Hawaii’s position on the plot?



4.32 Match the correlation. You are going to use the *Correlation and Regression* applet to make scatterplots with 10 points that have correlation close to 0.7. The lesson is that many patterns can have the same correlation. Always plot your data before you trust a correlation.

- (a) Stop after adding the first two points. What is the value of the correlation? Why does it have this value?
- (b) Make a lower-left to upper-right pattern of 10 points with correlation about $r = 0.7$. (You can drag points up or down to adjust r after you have 10 points.) Make a rough sketch of your scatterplot.

- (c) Make another scatterplot with 9 points in a vertical stack at the left of the plot. Add one point far to the right and move it until the correlation is close to 0.7. Make a rough sketch of your scatterplot.
- (d) Make yet another scatterplot with 10 points in a curved pattern that starts at the lower left, rises to the right, then falls again at the far right. Adjust the points up or down until you have a quite smooth curve with correlation close to 0.7. Make a rough sketch of this scatterplot also.

4.33 Correlation is not resistant. Go to the *Correlation and Regression* applet. Click on the scatterplot to create a group of 10 points in the lower-left corner of the scatterplot with a strong straight-line pattern (correlation about 0.9).



- (a) Add one point at the upper right that is in line with the first 10. How does the correlation change?
- (b) Drag this last point down until it is opposite the group of 10 points. How small can you make the correlation? Can you make the correlation negative? You see that a single outlier can greatly strengthen or weaken a correlation. Always plot your data to check for outlying points.

4.34 Brain size and IQ score. Do people with larger brains have higher IQ scores? A study looked at 40 volunteer subjects, 20 men and 20 women. Brain size was measured by magnetic resonance imaging. The EESEE story “Brain Size and Intelligence” gives the data. The MRI count is the number of “pixels” the brain covered in the image. IQ was measured by the Wechsler test that we met in Chapter 3.



- (a) Make a scatterplot of IQ score versus MRI count, using distinct symbols for men and women. In addition, find the correlation between IQ and MRI for all 40 subjects, for the men alone, and for the women alone.
- (b) Men are larger than women on the average, so they have larger brains. How is this size effect visible in your plot? Find the mean MRI count for men and women to verify the difference.
- (c) Your result in (b) suggests separating men and women in looking at the relationship between brain size and IQ. Use your work in (a) to comment on the nature and strength of this relationship for women and for men.

4.35 Mercury in bass. Mercury in fish can make eating the fish unsafe. The EESEE story “Mercury in Florida’s Bass” contains data from 53 lakes in Florida. We are interested in the relationship between the mercury concentration in the flesh of fish from a lake and various measures of water quality. EESEE asks you to examine several relationships. Start with this one: plot the standardized mercury concentration against the pH (the chemical measure of acidity versus alkalinity) for the 53 lakes. Give a brief verbal and numerical description of the relationship.



CHAPTER 5



(R.Lynn/Photo Researchers)

In this chapter we cover...

The least-squares regression line
Using technology
Facts about least-squares regression
Residuals
Influential observations
Cautions about correlation and regression
Association does not imply causation

Regression

Linear (straight-line) relationships between two quantitative variables are easy to understand and quite common. In Chapter 4, we found linear relationships in settings as varied as sparrowhawk colonies, icicle growth, and heating a home. Correlation measures the direction and strength of these relationships. When a scatterplot shows a linear relationship, we would like to summarize the overall pattern by drawing a line on the scatterplot. A *regression line* summarizes the relationship between two variables, but only in a specific setting: one of the variables helps explain or predict the other. That is, regression describes a relationship between an explanatory variable and a response variable.

REGRESSION LINE

A **regression line** is a straight line that describes how a response variable y changes as an explanatory variable x changes. We often use a regression line to predict the value of y for a given value of x .

EXAMPLE 5.1 Predicting new birds

We saw in Exercise 4.4 (page 83) that there is a linear relationship between the percent x of adult sparrowhawks that return to a colony from the previous year and the number y of new adult birds that join the colony. The scatterplot in Figure 5.1 displays this relationship.

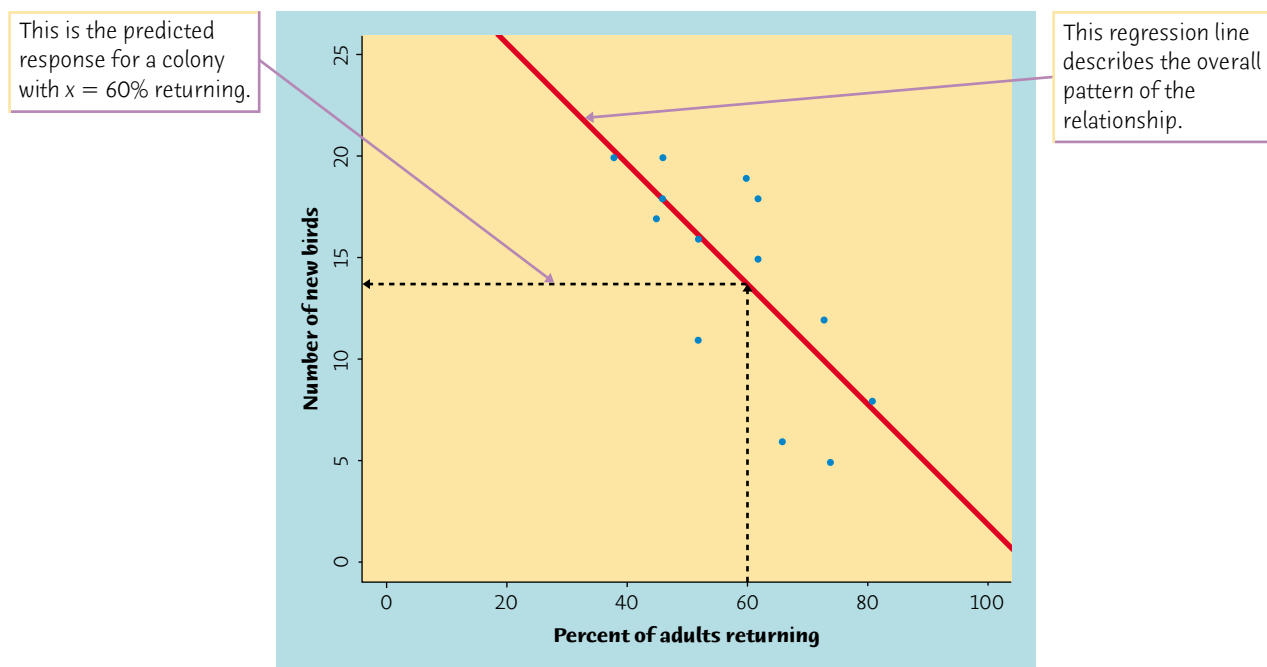


Figure 5.1 Data on 13 sparrowhawk colonies, with a regression line for predicting number of new birds from percent of returning birds. The dashed lines illustrate how to use the regression line to predict new birds in a colony with 60% returning.

The correlation is $r = -0.7485$, so the straight-line pattern is moderately strong. The line on the plot is a regression line that describes the overall pattern.

An ecologist wants to use the line, based on 13 colonies, to predict how many birds will join another colony, to which 60% of the adults from the previous year return. To **predict** new birds for 60% returning, first locate 60 on the x axis. Then go “up and over” as in the figure to find the y that corresponds to $x = 60$. It appears from the graph that we predict around 13 or 14 new birds.

prediction

The least-squares regression line

Different people will draw different lines by eye on a scatterplot. This is especially true when the points are widely scattered. We need a way to draw a regression line that doesn’t depend on our guess as to where the line should go. We will use the line to predict y from x , so the prediction errors we make are errors in y , the vertical direction in the scatterplot. If we predict 14 new birds for a colony with 60% returning birds and in fact 18 new birds join the colony, our prediction error is

$$\begin{aligned}\text{error} &= \text{observed } y - \text{predicted } y \\ &= 18 - 14 = 4\end{aligned}$$

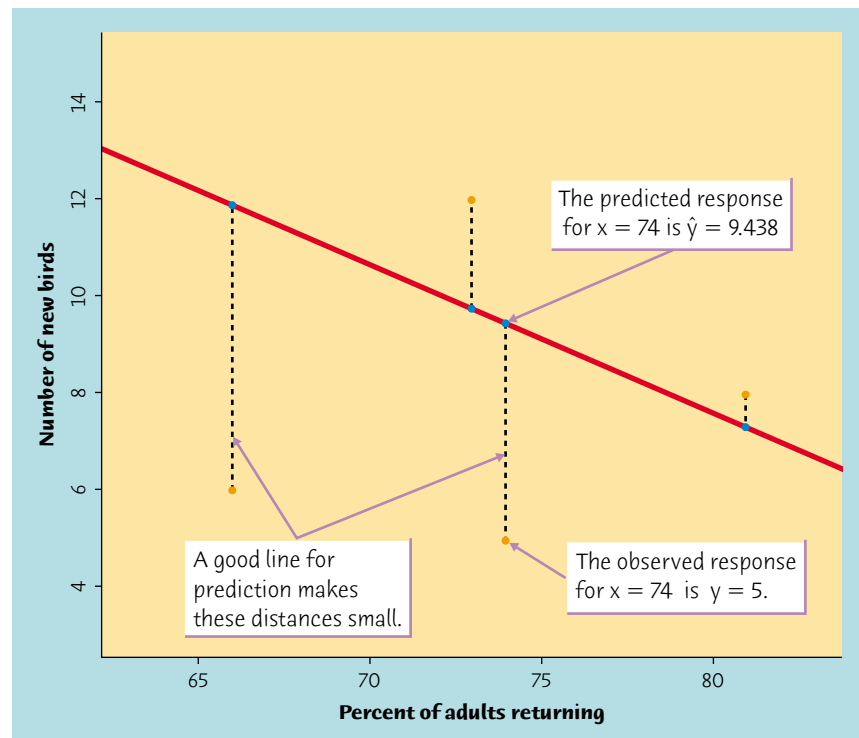


Figure 5.2 The least-squares idea. For each observation, find the vertical distance of each point on the scatterplot from a regression line. The least-squares regression line makes the sum of the squares of these distances as small as possible.

No line will pass exactly through all the points in the scatterplot. We want the *vertical* distances of the points from the line to be as small as possible. Figure 5.2 illustrates the idea. This plot shows four of the points from Figure 5.1, along with the line, on an expanded scale. The line passes above two of the points and below two of them. The vertical distances of the data points from the line appear as vertical line segments. There are many ways to make the collection of vertical distances “as small as possible.” The most common is the *least-squares* method.

LEAST-SQUARES REGRESSION LINE

The **least-squares regression line** of y on x is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.

One reason for the popularity of the least-squares regression line is that the problem of finding the line has a simple answer. We can give the recipe for the least-squares line in terms of the means and standard deviations of the two variables and their correlation.

EQUATION OF THE LEAST-SQUARES REGRESSION LINE

We have data on an explanatory variable x and a response variable y for n individuals. From the data, calculate the means $\bar{x}$ and $\bar{y}$ and the standard deviations s_x and s_y of the two variables, and their correlation r . The least-squares regression line is the line

$$\hat{y} = a + bx$$

with **slope**

$$b = r \frac{s_y}{s_x}$$

and **intercept**

$$a = \bar{y} - b\bar{x}$$

We write $\hat{y}$ (read “y hat”) in the equation of the regression line to emphasize that the line gives a *predicted* response $\hat{y}$ for any x . Because of the scatter of points about the line, the predicted response will usually not be exactly the same as the actually *observed* response y . In practice, you don’t need to calculate the means, standard deviations, and correlation first. Software or your calculator will give the slope b and intercept a of the least-squares line from keyed-in values of the variables x and y . You can then concentrate on understanding and using the regression line.

EXAMPLE 5.2 Using a regression line

The line in Figure 5.1 is in fact the least-squares regression line of new birds on percent of returning birds. Enter the data from Exercise 4.4 into your calculator and check that the equation of this line is

$$\hat{y} = 31.9343 - 0.3040x$$

The **slope** of a regression line is usually important for the interpretation of the data. The slope is the rate of change, the amount of change in $\hat{y}$ when x increases by 1. The slope $b = -0.3040$ in this example says that for each additional percent of last year’s birds that return we predict about 0.3 fewer new birds.

slope

The **intercept** of the regression line is the value of $\hat{y}$ when $x = 0$. Although we need the value of the intercept to draw the line, it is statistically meaningful only when x can actually take values close to zero. In our example, $x = 0$ means that a colony disappears because no birds return. The line predicts that on the average 31.9 new birds will appear. This isn’t meaningful because a colony disappearing is a different setting than a colony with returning birds.

intercept

The equation of the regression line makes **prediction** easy. Just substitute an x -value into the equation. To predict new birds when 60% return, substitute $x = 60$:

prediction

$$\begin{aligned}\hat{y} &= 31.9343 - (0.3040)(60) \\ &= 31.9343 - 18.24 = 13.69\end{aligned}$$

The actual number of new birds must be a whole number. Think of the prediction $\hat{y} = 13.69$ as an “on the average” value for many colonies with 60% returning birds.

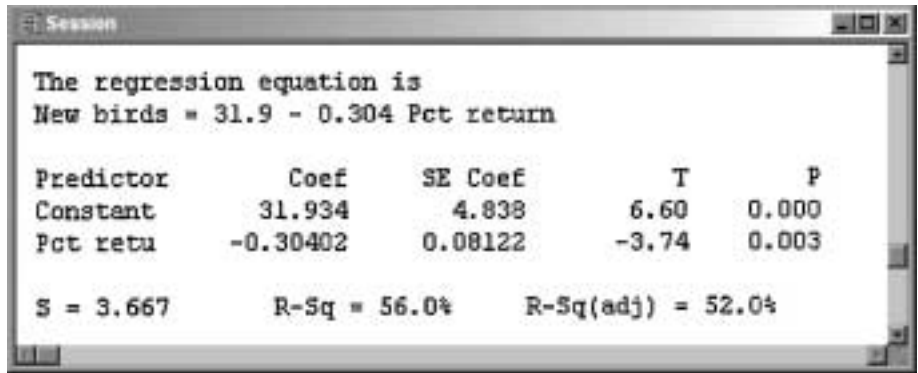
plotting a line

To **plot the line** on the scatterplot, use the equation to find $\hat{y}$ for two values of x , one near each end of the range of x in the data. Plot each $\hat{y}$ above its x and draw the line through the two points.

Using technology

Least-squares regression is one of the most common statistical procedures. Any technology you use for statistical calculations will give you the least-squares line and related information. Figure 5.3 displays the regression output for the

Minitab



Excel

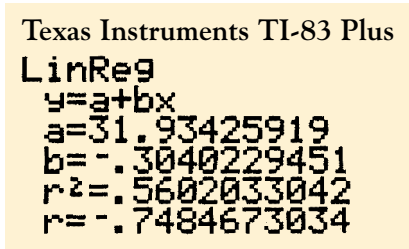
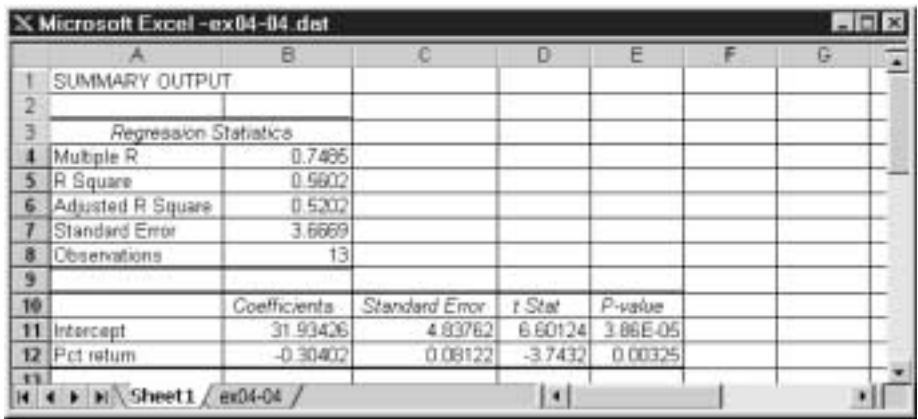


Figure 5.3 Least-squares regression for the sparrowhawk data. Output from statistical software, a spreadsheet, and a graphing calculator.

sparrowhawk data from a statistical software package, a spreadsheet program, and a graphing calculator. Each output records the slope and intercept of the least-squares line. The software also provides information that we do not yet need, although we will use much of it later. (In fact, we left out part of the Minitab and Excel outputs.) Be sure that you can locate the slope and intercept on all three outputs. Once you understand the statistical ideas, you can read and work with almost any software output.

APPLY YOUR KNOWLEDGE

- 5.1 Verify our claims.** Example 5.2 gives the equation of the regression line of new birds y on percent of returning birds x for the data in Exercise 4.4 as

$$\hat{y} = 31.9343 - 0.3040x$$

Enter the data from Exercise 4.4 into your calculator.

- (a) Use your calculator's regression function to find the equation of the least-squares regression line.
- (b) Use your calculator to find the mean and standard deviation of both x and y and their correlation r . Find the slope b and intercept a of the regression line from these, using the facts in the box Equation of the Least-Squares Regression Line. Verify that in both part (a) and part (b) you get the equation in Example 5.2. (Results may differ slightly because of rounding off.)

- 5.2 Penguins diving.** A study of king penguins looked for a relationship between how deep the penguins dive to seek food and how long they stay under water.¹ For all but the shallowest dives, there is a linear relationship that is different for different penguins. The study report gives a scatterplot for one penguin titled "The relation of dive duration (DD) to depth (D)." Duration DD is measured in minutes and depth D is in meters. The report then says, "The regression equation for this bird is: $DD = 2.69 + 0.0138D$."

- (a) What is the slope of the regression line? Explain in specific language what this slope says about this penguin's dives.
- (b) According to the regression line, how long does a typical dive to a depth of 200 meters last?
- (c) The dives varied from 40 meters to 300 meters in depth. Plot the regression line from $x = 40$ to $x = 300$.

- 5.3 Sports car gas mileage.** Table 1.2 (page 12) gives the city and highway gas mileages for two-seater cars. A scatterplot (Exercise 4.12) shows a strong positive linear relationship.

- (a) Find the least-squares regression line for predicting highway mileage from city mileage, using data from all 22 car models. Make a scatterplot and plot the regression line.



Regression toward the mean

To “regress” means to go backward. Why are statistical methods for predicting a response from an explanatory variable called “regression”? Sir Francis Galton (1822–1911), who was the first to apply regression to biological and psychological data, looked at examples such as the heights of children versus the heights of their parents. He found that the taller-than-average parents tended to have children who were also taller than average but not as tall as their parents. Galton called this fact “regression toward the mean,” and the name came to be applied to the statistical method.

- (b) What is the slope of the regression line? Explain in words what the slope says about gas mileage for two-seater cars.
- (c) Another two-seater is rated at 20 miles per gallon in the city. Predict its highway mileage.

Facts about least-squares regression

One reason for the popularity of least-squares regression lines is that they have many convenient special properties. Here are some facts about least-squares regression lines.

Fact 1. The distinction between explanatory and response variables is essential in regression. Least-squares regression looks at the distances of the data points from the line only in the y direction. If we reverse the roles of the two variables, we get a different least-squares regression line.

EXAMPLE 5.3 The expanding universe

Figure 5.4 is a scatterplot of data that played a central role in the discovery that the universe is expanding. They are the distances from earth of 24 spiral galaxies and the speed at which these galaxies are moving away from us, reported by the astronomer

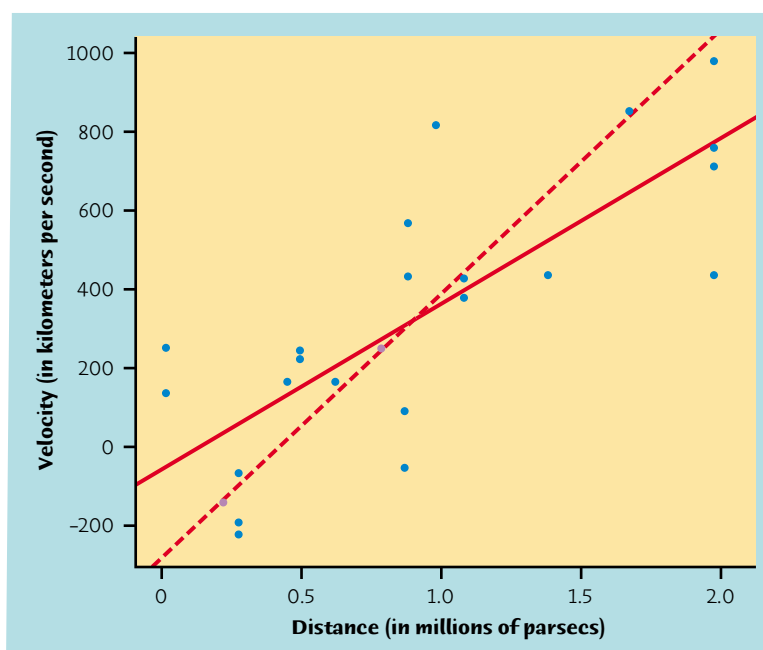


Figure 5.4 Scatterplot of Hubble’s data on the distance from earth of 24 galaxies and the velocity at which they are moving away from us. The two lines are the two least-squares regression lines: of velocity on distance (*solid*) and of distance on velocity (*dashed*).

Edwin Hubble in 1929.² There is a positive linear relationship, $r = 0.7842$, so that more distant galaxies are moving away more rapidly. Astronomers believe that there is in fact a perfect linear relationship, and that the scatter is caused by imperfect measurements.

The two lines on the plot are the two least-squares regression lines. The regression line of velocity on distance is solid. The regression line of distance on velocity is dashed. *Regression of velocity on distance and regression of distance on velocity give different lines.* In the regression setting you must know clearly which variable is explanatory.

Fact 2. There is a close connection between correlation and the slope of the least-squares line. The slope is

$$b = r \frac{s_y}{s_x}$$

This equation says that along the regression line, **a change of one standard deviation in x corresponds to a change of r standard deviations in y .** When the variables are perfectly correlated ($r = 1$ or $r = -1$), the change in the predicted response $\hat{y}$ is the same (in standard deviation units) as the change in x . Otherwise, because $-1 \leq r \leq 1$, the change in $\hat{y}$ is less than the change in x . As the correlation grows less strong, the prediction $\hat{y}$ moves less in response to changes in x .

Fact 3. The least-squares regression line always passes through the point $(\bar{x}, \bar{y})$ on the graph of y against x . So the least-squares regression line of y on x is the line with slope rs_y/s_x that passes through the point $(\bar{x}, \bar{y})$.

Fact 4. The correlation r describes the strength of a straight-line relationship. In the regression setting, this description takes a specific form: **the square of the correlation, r^2 , is the fraction of the variation in the values of y that is explained by the least-squares regression of y on x .**

The idea is that when there is a linear relationship, some of the variation in y is accounted for by the fact that as x changes it pulls y along with it. Look again at Figure 5.1 on page 105. The number of new birds joining a colony ranges from 5 to 20. Some of this variation in the response y is explained by the fact that the percent x of returning birds varies from 38% to 81%. As x moves from 38% to 81%, it pulls y with it along the line. You would guess a smaller number of new birds for a colony with 80% returning than for a colony with 40% returning. But there is also quite a bit of scatter above and below the line, variation that isn't explained by the straight-line relationship between x and y .

Although we won't do the algebra, it is possible to break the total variation in the observed values of y into two parts. One part is the variation we expect as x moves and $\hat{y}$ moves with it along the regression line. The other measures the variation of the data points about the line. The squared correlation r^2 is

the first of these as a fraction of the whole:

$$r^2 = \frac{\text{variation in } \hat{y} \text{ as } x \text{ pulls it along the line}}{\text{total variation in observed values of } y}$$

EXAMPLE 5.4 Using r^2

In Figure 5.1, $r = -0.7485$ and $r^2 = 0.5603$. About 56% of the variation in new birds is accounted for by the linear relationship with percent returning. The other 44% is individual variation among colonies that is not explained by the linear relationship.

Figure 4.2 (page 85) shows a stronger linear relationship in which the points are more tightly concentrated along a line. Here, $r = -0.9124$ and $r^2 = 0.8325$. More than 83% of the variation in carnivore abundance is explained by regression on body mass. Only 17% is variation among species with the same mass.

When you report a regression, give r^2 as a measure of how successful the regression was in explaining the response. All the outputs in Figure 5.3 include r^2 , either in decimal form or as a percent. When you see a correlation, square it to get a better feel for the strength of the association. Perfect correlation ($r = -1$ or $r = 1$) means the points lie exactly on a line. Then $r^2 = 1$ and all of the variation in one variable is accounted for by the linear relationship with the other variable. If $r = -0.7$ or $r = 0.7$, $r^2 = 0.49$ and about half the variation is accounted for by the linear relationship. In the r^2 scale, correlation ± 0.7 is about halfway between 0 and ± 1 .

Facts 2, 3, and 4 are special properties of least-squares regression. They are not true for other methods of fitting a line to data.

APPLY YOUR KNOWLEDGE

5.4 Growing corn. Exercise 4.25 (page 99) gives data from an agricultural experiment. The purpose of the study was to see how the yield of corn changes as we change the planting rate (plants per acre).

- Make a scatterplot of the data. (Use a scale of yields from 100 to 200 bushels per acre.) Find the least-squares regression line for predicting yield from planting rate and add this line to your plot. Why should we *not* use regression for prediction in this setting?
- What is r^2 ? What does this value say about the success of the regression in predicting yield?
- Even regression lines that make no practical sense obey Facts 1 to 4. Use the equation of the regression line you found in (a) to show that when x is the mean planting rate, the predicted yield $\hat{y}$ is the mean of the observed yields.

5.5 Sports car gas mileage. In Exercise 5.3 you found the least-squares regression line for predicting highway mileage from city mileage for the

22 two-seater car models in Table 1.2. Find the mean city mileage and mean highway mileage for these cars. Use your regression line to predict the highway mileage for a car with city mileage equal to the mean for the group. Explain why you knew the answer before doing the prediction.

- 5.6 Comparing regressions.** What is the value of r^2 for predicting highway from city mileage in Exercise 5.5? What value did you find for predicting corn yield from planting rate in Exercise 5.4? Explain in simple language why if we knew only these two r^2 -values, we would expect predictions using the regression line to be more satisfactory for gas mileage than for corn yield.

Residuals

One of the first principles of data analysis is to look for an overall pattern and also for striking deviations from the pattern. A regression line describes the overall pattern of a linear relationship between an explanatory variable and a response variable. We see deviations from this pattern by looking at the scatter of the data points about the regression line. The vertical distances from the points to the least-squares regression line are as small as possible, in the sense that they have the smallest possible sum of squares. Because they represent “left-over” variation in the response after fitting the regression line, these distances are called *residuals*.

RESIDUALS

A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line. That is,

$$\begin{aligned}\text{residual} &= \text{observed } y - \text{predicted } y \\ &= y - \hat{y}\end{aligned}$$

EXAMPLE 5.5 Predicting mental ability

Does the age at which a child begins to talk predict later score on a test of mental ability? A study of the development of young children recorded the age in months at which each of 21 children spoke their first word and their Gesell Adaptive Score, the result of an aptitude test taken much later. The data appear in Table 5.1.³

Figure 5.5 is a scatterplot, with age at first word as the explanatory variable x and Gesell score as the response variable y . Children 3 and 13, and also Children 16 and 21, have identical values of both variables. We use a different plotting symbol to show that one point stands for two individuals. The plot shows a negative association. That is, children who begin to speak later tend to have lower test scores than early talkers. The overall pattern is moderately linear. The correlation describes both the direction and the strength of the linear relationship. It is $r = -0.640$.

The line on the plot is the least-squares regression line of Gesell score on age at first word. Its equation is

$$\hat{y} = 109.8738 - 1.1270x$$

TABLE 5.1 Age at first word and Gesell score

Child	Age	Score	Child	Age	Score
1	15	95	11	7	113
2	26	71	12	9	96
3	10	83	13	10	83
4	9	91	14	11	84
5	15	102	15	11	102
6	20	87	16	10	100
7	18	93	17	12	105
8	11	100	18	42	57
9	8	104	19	17	121
10	20	94	20	11	86
			21	10	100

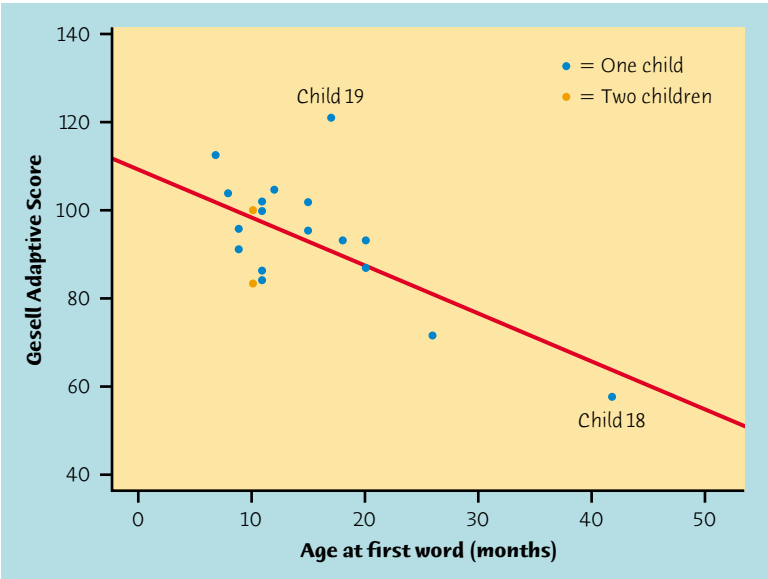


Figure 5.5 Scatterplot of Gesell Adaptive Score versus the age at first word for 21 children, from Table 5.1. The line is the least-squares regression line for predicting Gesell score from age at first word.

For Child 1, who first spoke at 15 months, we predict the score

$$\hat{y} = 109.8738 - (1.1270)(15) = 92.97$$

This child's actual score was 95. The residual is

$$\begin{aligned} \text{residual} &= \text{observed } y - \text{predicted } y \\ &= 95 - 92.97 = 2.03 \end{aligned}$$

The residual is positive because the data point lies above the line.

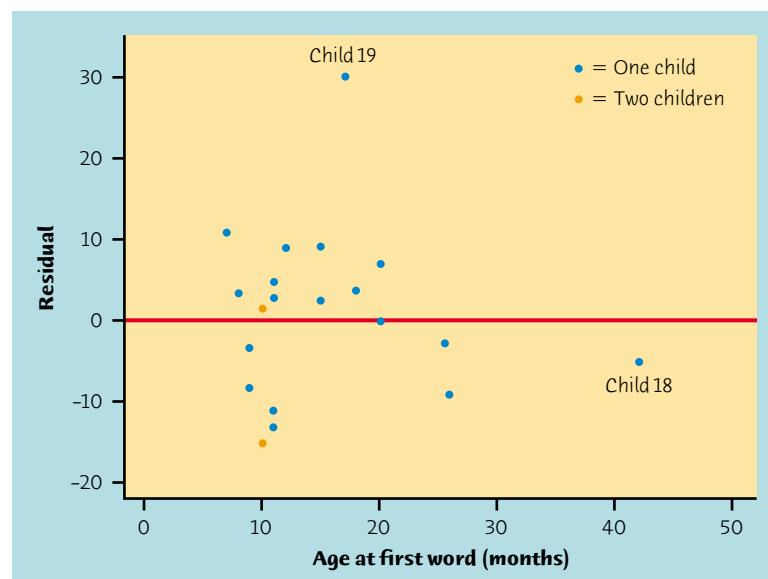


Figure 5.6 Residual plot for the regression of Gesell score on age at first word. Child 19 is an outlier. Child 18 is an influential observation that does not have a large residual.

There is a residual for each data point. Finding the residuals is a bit unpleasant because you must first find the predicted response for every x . Software or a graphing calculator gives you the residuals all at once. Here are the 21 residuals for the Gesell data, from software:

```
residuals:
2.0310 -9.5721 -15.6040 -8.7309  9.0310  -0.3341  3.4120
2.5230  3.1421  6.6659 11.0151 -3.7309 -15.6040 -13.4770
4.5230  1.3960  8.6500 -5.5403 30.2850 -11.4770  1.3960
```

Because the residuals show how far the data fall from our regression line, examining the residuals helps assess how well the line describes the data. Although residuals can be calculated from any model fitted to the data, the residuals from the least-squares line have a special property: **the mean of the least-squares residuals is always zero.**

Compare the scatterplot in Figure 5.5 with the *residual plot* for the same data in Figure 5.6. The horizontal line at zero in Figure 5.6 helps orient us. It corresponds to the regression line in Figure 5.5.

RESIDUAL PLOTS

A **residual plot** is a scatterplot of the regression residuals against the explanatory variable. Residual plots help us assess the fit of a regression line.

By in effect turning the regression line horizontal, a residual plot magnifies the deviations of the points from the line and makes it easier to see unusual observations and patterns.

APPLY YOUR KNOWLEDGE

5.7 Does fast driving waste fuel? Exercise 4.6 (page 86) gives data on the fuel consumption y of a car at various speeds x . Fuel consumption is measured in liters of gasoline per 100 kilometers driven and speed is measured in kilometers per hour. Software tells us that the equation of the least-squares regression line is

$$\hat{y} = 11.058 - 0.01466x$$

The residuals, in the same order as the observations, are

10.09	2.24	−0.62	−2.47	−3.33	−4.28	−3.73	−2.94
−2.17	−1.32	−0.42	0.57	1.64	2.76	3.97	

- Make a scatterplot of the observations and draw the regression line on your plot.
- Would you use the regression line to predict y from x ? Explain your answer.
- Check that the residuals have sum zero (up to roundoff error).
- Make a plot of the residuals against the values of x . Draw a horizontal line at height zero on your plot. Notice that the residuals show the same pattern about this line as the data points show about the regression line in the scatterplot in (a).

Influential observations

Figures 5.5 and 5.6 show two unusual observations. Children 18 and 19 are unusual in different ways. Child 19 lies far from the regression line. Child 18 is close to the line but far out in the x direction. Child 19 is an *outlier in the y direction*, with a Gesell score so high that we should check for a mistake in recording it. In fact, the score is correct.

Child 18 is an *outlier in the x direction*. This child began to speak much later than any of the other children. Because of its extreme position on the age scale, this point has a strong influence on the position of the regression line. Figure 5.7 adds a second regression line, calculated after leaving out Child 18. You can see that this one point moves the line quite a bit. Least-squares lines make the sum of squares of the vertical distances to the points as small as possible. A point that is extreme in the x direction with no other points near it pulls the line toward itself. We call such points *influential*.

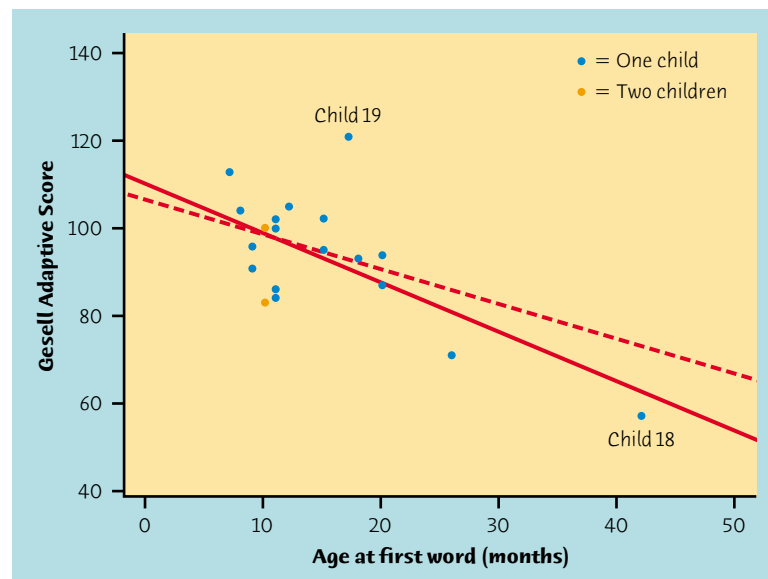


Figure 5.7 Two least-squares regression lines of Gesell score on age at first word. The solid line is calculated from all the data. The dashed line was calculated leaving out Child 18. Child 18 is an influential observation because leaving out this point moves the regression line quite a bit.

OUTLIERS AND INFLUENTIAL OBSERVATIONS IN REGRESSION

An **outlier** is an observation that lies outside the overall pattern of the other observations. Points that are outliers in the y direction of a scatterplot have large regression residuals, but other outliers need not have large residuals.

An observation is **influential** for a statistical calculation if removing it would markedly change the result of the calculation. Points that are outliers in the x direction of a scatterplot are often influential for the least-squares regression line.

We did not need the distinction between outliers and influential observations in Chapter 2. A single large salary that pulls up the mean salary $\bar{x}$ for a group of workers is an outlier because it lies far above the other salaries. It is also influential, because the mean changes when it is removed. In the regression setting, however, not all outliers are influential. The least-squares regression line is most likely to be heavily influenced by observations that are outliers in the x direction. The scatterplot will alert you to observations that are extreme in x and may therefore be influential. The surest way to verify that a point is influential is to find the regression line both with and without the suspect point, as in Figure 5.7. If the line moves more than a small amount when the point



is deleted, the point is influential. The *Correlation and Regression* applet allows you to move points and watch how they influence the least-squares line.

EXAMPLE 5.6 An influential observation

The strong influence of Child 18 makes the original regression of Gesell score on age at first word misleading. The original data have $r^2 = 0.41$. That is, the age at which a child begins to talk explains 41% of the variation on a later test of mental ability. This relationship is strong enough to be interesting to parents. If we leave out Child 18, r^2 drops to only 11%. The apparent strength of the association was largely due to a single influential observation.

What should the child development researcher do? She must decide whether Child 18 was so slow to speak that this individual should not be allowed to influence the analysis. If she excludes Child 18, much of the evidence for a connection between the age at which a child begins to talk and later ability score vanishes. If she keeps Child 18, she needs data on other children who were also slow to begin talking, so that the analysis no longer depends so heavily on just one child.

APPLY YOUR KNOWLEDGE

5.8 Influential or not? We have seen that Child 18 in the Gesell data in Table 5.1 (page 114) is an influential observation. Now we will examine the effect of Child 19, who is also an outlier in Figure 5.5.

- (a) Find the least-squares regression line of Gesell score on age at first word, leaving out Child 19. Example 5.5 gives the regression line from all the children. Plot both lines on the same graph. (You do not have to make a scatterplot of all the points; just plot the two lines.) Would you call Child 19 very influential? Why?
- (b) For all children, $r^2 = 0.41$. How does removing Child 19 change the r^2 for this regression? Explain why r^2 changes in this direction when you drop Child 19.

5.9 Sports car gas mileage. The data on gas mileage of two-seater cars (Table 1.2, page 12) contain an outlier, the Honda Insight. When we predict highway mileage from city mileage, this point is an outlier in *both* the x and y directions. We wonder if it influences the least-squares line.

- (a) Make a scatterplot and draw (again) the least-squares line from all 22 car models.
- (b) Find the least-squares line when the Insight is left out of the calculation and draw this line on your plot.
- (c) Influence is a matter of degree, not a yes-or-no question. Use both regression lines to predict highway mileages for city mileages of 10, 20, and 25 MPG. (These city mileage values span the range of car models other than the Insight.) Do you think the Insight changes the predictions enough to be important to a car buyer?

Cautions about correlation and regression

Correlation and regression are powerful tools for describing the relationship between two variables. When you use these tools, you must be aware of their limitations, beginning with the fact that **correlation and regression describe only linear relationships**. Also remember that **the correlation r and the least-squares regression line are not resistant**. One influential observation or incorrectly entered data point can greatly change these measures. **Always plot your data** before interpreting regression or correlation. Here are some other cautions to keep in mind when you apply correlation and regression or read accounts of their use.

Beware extrapolation. Suppose that you have data on a child's growth between 3 and 8 years of age. You find a strong linear relationship between age x and height y . If you fit a regression line to these data and use it to predict height at age 25 years, you will predict that the child will be 8 feet tall. Growth slows down and then stops at maturity, so extending the straight line to adult ages is foolish. Few relationships are linear for all values of x . So don't stray far from the range of x that actually appears in your data.

EXTRAPOLATION

Extrapolation is the use of a regression line for prediction far outside the range of values of the explanatory variable x that you used to obtain the line. Such predictions are often not accurate.

Beware the lurking variable. Correlation and regression describe the relationship between two variables. Often the relationship between two variables is strongly influenced by other variables. More advanced statistical methods allow the study of many variables together, so that we can take other variables into account. Sometimes, however, the relationship between two variables is influenced by other variables that we did not measure or even think about. Because these variables are lurking in the background, we call them *lurking variables*.

LURKING VARIABLE

A **lurking variable** is a variable that has an important effect on the relationship among the variables in a study but is not included among the variables studied.

You should always think about possible lurking variables before you draw conclusions based on correlation or regression.



Do left-handers die early?

Yes, said a study of 1000 deaths in California. Left-handed people died at an average age of 66 years; right-handers, at 75 years of age. Should left-handed people fear an early death? No—the lurking variable has struck again. Older people grew up in an era when many natural left-handers were forced to use their right hands. So right-handers are more common among older people, and left-handers are more common among the young. When we look at deaths, the left-handers who die are younger on the average because left-handers in general are younger. Mystery solved.

EXAMPLE 5.7 Magic Mozart?

The Kalamazoo (Michigan) Symphony once advertised a “Mozart for Minors” program with this statement: “Question: Which students scored 51 points higher in verbal skills and 39 points higher in math? Answer: Students who had experience in music.”⁴

We could as well answer “Children who played soccer.” Why? Children with prosperous and well-educated parents are more likely than poorer children to have experience with music and also to play soccer. They are also likely to attend good schools, get good health care, and be encouraged to study hard. These advantages lead to high test scores. Experience with music and soccer are correlated with high scores just because they go along with the other advantages of having prosperous and educated parents.

APPLY YOUR KNOWLEDGE

5.10 The declining farm population. The number of people living on American farms has declined steadily during this century. Here are data on the farm population (millions of persons) from 1935 to 1980:

Year	1935	1940	1945	1950	1955	1960	1965	1970	1975	1980
Population	32.1	30.5	24.4	23.0	19.1	15.6	12.4	9.7	8.9	7.2

- (a) Make a scatterplot of these data and find the least-squares regression line of farm population on year.
- (b) According to the regression line, how much did the farm population decline each year on the average during this period? What percent of the observed variation in farm population is accounted for by linear change over time?
- (c) Use the regression equation to predict the number of people living on farms in 1990. Is this result reasonable? Why?

5.11 Is math the key to success in college? A College Board study of 15,941 high school graduates found a strong correlation between how much math minority students took in high school and their later success in college. News articles quoted the head of the College Board as saying that “math is the gatekeeper for success in college.”⁵ Maybe so, but we should also think about lurking variables. What might lead minority students to take more or fewer high school math courses? Would these same factors influence success in college?

Association does not imply causation

Thinking about lurking variables leads to the most important caution about correlation and regression. When we study the relationship between two variables, we often hope to show that changes in the explanatory variable *cause*

changes in the response variable. A strong association between two variables is not enough to draw conclusions about cause and effect. Sometimes an observed association really does reflect cause and effect. The Sanchez household uses more natural gas in colder months because cold weather requires burning more gas to stay warm. In other cases, an association is explained by lurking variables, and the conclusion that x causes y is either wrong or not proved.

EXAMPLE 5.8 Does TV make you live longer?

Measure the number of television sets per person x and the average life expectancy y for the world's nations. There is a high positive correlation: nations with many TV sets have higher life expectancies.

The basic meaning of causation is that by changing x we can bring about a change in y . Could we lengthen the lives of people in Rwanda by shipping them TV sets? No. Rich nations have more TV sets than poor nations. Rich nations also have longer life expectancies because they offer better nutrition, clean water, and better health care. There is no cause-and-effect tie between TV sets and length of life.

Correlations such as that in Example 5.8 are sometimes called “nonsense correlations.” The correlation is real. What is nonsense is the conclusion that changing one of the variables causes changes in the other. A lurking variable—such as national wealth in Example 5.8—that influences both x and y can create a high correlation even though there is no direct connection between x and y .

ASSOCIATION DOES NOT IMPLY CAUSATION

An association between an explanatory variable x and a response variable y , even if it is very strong, is not by itself good evidence that changes in x actually cause changes in y .

EXAMPLE 5.9 Obesity in mothers and daughters

Obese parents tend to have obese children. The results of a study of Mexican American girls aged 9 to 12 years are typical. The investigators measured body mass index (BMI), a measure of weight relative to height, for both the girls and their mothers. People with high BMI are overweight or obese. The correlation between the BMI of daughters and the BMI of their mothers was $r = 0.506$.⁶

Body type is in part determined by heredity. Daughters inherit half their genes from their mothers. There is therefore a direct causal link between the BMI of mothers and daughters. But it may be that mothers who are overweight also set an example of little exercise, poor eating habits, and lots of television. Their daughters pick up these habits to some extent, so the influence of heredity is mixed up with influences from the girls' environment. Both contribute to the mother-daughter correlation.

The lesson of Example 5.9 is more subtle than just “association does not imply causation.” Even when direct causation is present, it may not be the whole

experiment

explanation for a correlation. You must still worry about lurking variables. Careful statistical studies try to anticipate lurking variables and measure them so that they are no longer “lurking.” The mother-daughter study did measure TV viewing, exercise, and diet. Elaborate statistical analysis can remove the effects of these variables to come closer to the direct effect of mother’s BMI on daughter’s BMI. This remains a second-best approach to causation. The best way to get good evidence that x causes y is to do an **experiment** in which we change x and keep lurking variables under control. We will discuss experiments in Chapter 8.

When experiments cannot be done, finding the explanation for an observed association is often difficult and controversial. Many of the sharpest disputes in which statistics plays a role involve questions of causation that cannot be settled by experiment. Do gun control laws reduce violent crime? Does using cell phones cause brain tumors? Has increased free trade widened the gap between the incomes of more educated and less educated American workers? All of these questions have become public issues. All concern associations among variables. And all have this in common: they try to pinpoint cause and effect in a setting involving complex relations among many interacting variables.

EXAMPLE 5.10 Does smoking cause lung cancer?

Despite the difficulties, it is sometimes possible to build a strong case for causation in the absence of experiments. The evidence that smoking causes lung cancer is about as strong as nonexperimental evidence can be.

Doctors had long observed that most lung cancer patients were smokers. Comparison of smokers and “similar” nonsmokers showed a very strong association between smoking and death from lung cancer. Could the association be explained by lurking variables? Might there be, for example, a genetic factor that predisposes people both to nicotine addiction and to lung cancer? Smoking and lung cancer would then be positively associated even if smoking had no direct effect on the lungs. How were these objections overcome?

Let’s answer this question in general terms: What are the criteria for establishing causation when we cannot do an experiment?

- *The association is strong.* The association between smoking and lung cancer is very strong.
- *The association is consistent.* Many studies of different kinds of people in many countries link smoking to lung cancer. That reduces the chance that a lurking variable specific to one group or one study explains the association.
- *Higher doses are associated with stronger responses.* People who smoke more cigarettes per day or who smoke over a longer period get lung cancer more often. People who stop smoking reduce their risk.
- *The alleged cause precedes the effect in time.* Lung cancer develops after years of smoking. The number of men dying of lung cancer rose as smoking became more common, with a lag of about 30 years. Lung cancer

kills more men than any other form of cancer. Lung cancer was rare among women until women began to smoke. Lung cancer in women rose along with smoking, again with a lag of about 30 years, and has now passed breast cancer as the leading cause of cancer death among women.

- *The alleged cause is plausible.* Experiments with animals show that tars from cigarette smoke do cause cancer.

Medical authorities do not hesitate to say that smoking causes lung cancer. The U.S. Surgeon General has long stated that cigarette smoking is “the largest avoidable cause of death and disability in the United States.”⁷ The evidence for causation is overwhelming—but it is not as strong as the evidence provided by well-designed experiments.

APPLY YOUR KNOWLEDGE

- 5.12 Education and income.** There is a strong positive association between the education and income of adults. For example, the Census Bureau reports that the median income of people aged 25 and over increases from \$15,800 for those with less than a ninth-grade education, to \$24,656 for high school graduates, to \$40,939 for holders of a bachelor’s degree, and on up for yet more education. In part, this association reflects causation—education helps people qualify for better jobs. Suggest several lurking variables (ask yourself what kinds of people tend to get good educations) that also contribute.
- 5.13 How’s your self-esteem?** People who do well tend to feel good about themselves. Perhaps helping people feel good about themselves will help them do better in school and life. Raising self-esteem became for a time a goal in many schools. California even created a state commission to advance the cause. Can you think of explanations for the association between high self-esteem and good school performance other than “Self-esteem causes better work in school”?
- 5.14 Are big hospitals bad for you?** A study shows that there is a positive correlation between the size of a hospital (measured by its number of beds x) and the median number of days y that patients remain in the hospital. Does this mean that you can shorten a hospital stay by choosing a small hospital? Why?

Chapter 5 SUMMARY

A **regression line** is a straight line that describes how a response variable y changes as an explanatory variable x changes. You can use a regression line to **predict** the value of y for any value of x by substituting this x into the equation of the line.

The **slope** b of a regression line $\hat{y} = a + bx$ is the rate at which the predicted response $\hat{y}$ changes along the line as the explanatory variable x changes. Specifically, b is the change in $\hat{y}$ when x increases by 1.

The **intercept** a of a regression line $\hat{y} = a + bx$ is the predicted response $\hat{y}$ when the explanatory variable $x = 0$. This prediction is of no statistical use unless x can actually take values near 0.

The most common method of fitting a line to a scatterplot is least squares. The **least-squares regression line** is the straight line $\hat{y} = a + bx$ that minimizes the sum of the squares of the vertical distances of the observed points from the line.

The least-squares regression line of y on x is the line with slope rs_y/s_x and intercept $a = \bar{y} - b\bar{x}$. This line always passes through the point $(\bar{x}, \bar{y})$.

Correlation and regression are closely connected. The correlation r is the slope of the least-squares regression line when we measure both x and y in standardized units. The **square of the correlation** r^2 is the fraction of the variance of one variable that is explained by least-squares regression on the other variable.

Correlation and regression must be **interpreted with caution**. **Plot the data** to be sure the relationship is roughly linear and to detect outliers and influential observations. A plot of the **residuals** makes these effects easier to see.

Look for **influential observations**, individual points that substantially change the regression line. Influential observations are often outliers in the x direction.

Avoid **extrapolation**, the use of a regression line for prediction for values of the explanatory variable far outside the range of the data from which the line was calculated.

Lurking variables that you did not measure may explain the relations between the variables you did measure. Correlation and regression can be misleading if you ignore important lurking variables.

Most of all, be careful not to conclude that there is a cause-and-effect relationship between two variables just because they are strongly associated. **High correlation does not imply causation**. The best evidence that an association is due to causation comes from an **experiment** in which the explanatory variable is directly changed and other influences on the response are controlled.

Chapter 5 EXERCISES

5.15 Sisters and brothers. How strongly do physical characteristics of sisters and brothers correlate? Here are data on the heights (in inches) of 11 adult pairs:⁸

Brother	71	68	66	67	70	71	70	73	72	65	66
Sister	69	64	65	63	65	62	65	64	66	59	62

- (a) Verify using your calculator or software that the least-squares line for predicting sister's height from brother's height is $\hat{y} = -6.356$

+ 1.048x. What is the correlation between sister's height and brother's height?

(b) Damien is 70 inches tall. Predict the height of his sister Tonya.

5.16 Husbands and wives. The mean height of American women in their twenties is about 64 inches, and the standard deviation is about 2.7 inches. The mean height of men the same age is about 69.3 inches, with standard deviation about 2.8 inches. If the correlation between the heights of husbands and wives is about $r = 0.5$, what is the slope of the regression line of the husband's height on the wife's height in young couples? Draw a graph of this regression line. Predict the height of the husband of a woman who is 67 inches tall.

5.17 Measuring water quality. Biochemical oxygen demand (BOD) measures organic pollutants in water by measuring the amount of oxygen consumed by microorganisms that break down these compounds. BOD is hard to measure accurately. Total organic carbon (TOC) is easy to measure, so it is common to measure TOC and use regression to predict BOD. A typical regression equation for water entering a municipal treatment plant is⁹

$$\text{BOD} = -55.43 + 1.507 \text{ TOC}$$

Both BOD and TOC are measured in milligrams per liter of water.

- What does the slope of this line say about the relationship between BOD and TOC?
- What is the predicted BOD when $\text{TOC} = 0$? Values of BOD less than 0 are impossible. Why does the prediction give an impossible value?

5.18 IQ and school GPA. Figure 4.6 (page 95) plots school grade point average (GPA) against IQ test score for 78 seventh-grade students. Calculation shows that the mean and standard deviation of the IQ scores are

$$\bar{x} = 108.9 \quad s_x = 13.17$$

For the grade point averages,

$$\bar{y} = 7.447 \quad s_y = 2.10$$

The correlation between IQ and GPA is $r = 0.6337$.

- Find the equation of the least-squares line for predicting GPA from IQ.
- What percent of the observed variation in these students' GPAs can be explained by the linear relationship between GPA and IQ?
- One student has an IQ of 103 but a very low GPA of 0.53. What is the predicted GPA for a student with $\text{IQ} = 103$? What is the residual for this particular student?



(Superstock/Superstock/PictureQuest)

5.19 A growing child. Sarah’s parents are concerned that she seems short for her age. Their doctor has the following record of Sarah’s height:

Age (months)	36	48	51	54	57	60
Height (cm)	86	90	91	93	94	95

- (a) Make a scatterplot of these data. Note the strong linear pattern.
- (b) Using your calculator, find the equation of the least-squares regression line of height on age.
- (c) Predict Sarah’s height at 40 months and at 60 months. Use your results to draw the regression line on your scatterplot.
- (d) What is Sarah’s rate of growth, in centimeters per month? Normally growing girls gain about 6 cm in height between ages 4 (48 months) and 5 (60 months). What rate of growth is this in centimeters per month? Is Sarah growing more slowly than normal?
- 5.20 Heating a home.** Exercise 4.16 (page 96) gives data on degree-days and natural gas consumed by the Sanchez home for 16 consecutive months. There is a very strong linear relationship. Mr. Sanchez asks, “If a month averages 20 degree-days per day (that’s 45° F), how much gas will we use?” Use your calculator or software to find the least-squares regression line and answer his question.
- 5.21 A nonsense prediction.** Use the least-squares regression line for the data in Exercise 5.19 to predict Sarah’s height at age 40 years (480 months). Your prediction is in centimeters. Convert it to inches using the fact that a centimeter is 0.3937 inch. The data have r^2 almost 0.99. Why is the prediction clearly silly?
- 5.22 Merlins breeding.** Exercise 4.20 (page 97) gives data on the number of breeding pairs of merlins in an isolated area in each of nine years and the percent of males who returned the next year. The data show that the percent returning is lower after successful breeding seasons and that the relationship is roughly linear. Use your calculator or software to find the least-squares regression line and predict the percent of returning males after a season with 30 breeding pairs.
- 5.23 Keeping water clean.** Keeping water supplies clean requires regular measurement of levels of pollutants. The measurements are indirect—a typical analysis involves forming a dye by a chemical reaction with the dissolved pollutant, then passing light through the solution and measuring its “absorbence.” To calibrate such measurements, the laboratory measures known standard solutions and uses regression to relate absorbence to pollutant concentration. This is usually done every day. Here is one series of data on the absorbence for different levels of nitrates. Nitrates are measured in milligrams per liter of water.¹⁰

Nitrates	50	50	100	200	400	800	1200	1600	2000	2000
Absorbence	7.0	7.5	12.8	24.0	47.0	93.0	138.0	183.0	230.0	226.0

- (a) Chemical theory says that these data should lie on a straight line. If the correlation is not at least 0.997, something went wrong and the calibration procedure is repeated. Plot the data and find the correlation. Must the calibration be done again?
- (b) What is the equation of the least-squares line for predicting absorbence from concentration? If the lab analyzed a specimen with 500 milligrams of nitrates per liter, what do you expect the absorbence to be? Based on your plot and the correlation, do you expect your predicted absorbence to be very accurate?

5.24 Comparing regressions. What are the correlations between the explanatory and response variables in Exercises 5.20 and 5.22? What does r^2 say about the two regressions? Which of the two predictions do you expect to be more accurate? Explain why.

5.25 Is wine good for your heart? Table 4.3 (page 101) gives data on wine consumption and heart disease death rates in 19 countries. A scatterplot (Exercise 4.27) shows a moderately strong relationship.

- (a) The correlation for these variables is $r = -0.843$. What does a negative correlation say about wine consumption and heart disease deaths? About what percent of the variation among countries in heart disease death rates is explained by the straight-line relationship with wine consumption?
- (b) The least-squares regression line for predicting heart disease death rate from wine consumption is

$$\hat{y} = 260.56 - 22.969x$$

Use this equation to predict the heart disease death rate in another country where adults average 4 liters of alcohol from wine each year.

- (c) The correlation in (a) and the slope of the least-squares line in (b) are both negative. Is it possible for these two quantities to have opposite signs? Explain your answer.

5.26 Always plot your data! Table 5.2 presents four sets of data prepared by the statistician Frank Anscombe to illustrate the dangers of calculating without first plotting the data.¹¹

- (a) Without making scatterplots, find the correlation and the least-squares regression line for all four data sets. What do you notice? Use the regression line to predict y for $x = 10$.
- (b) Make a scatterplot for each of the data sets and add the regression line to each plot.



(Digital Vision/Getty Images)

TABLE 5.2 Four data sets for exploring correlation and regression

Data Set A											
x	10	8	13	9	11	14	6	4	12	7	5
y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68

Data Set B											
x	10	8	13	9	11	14	6	4	12	7	5
y	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74

Data Set C											
x	10	8	13	9	11	14	6	4	12	7	5
y	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73

Data Set D											
x	8	8	8	8	8	8	8	8	8	8	19
y	6.58	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.50

(c) In which of the four cases would you be willing to use the regression line to describe the dependence of y on x ? Explain your answer in each case.

5.27 Lots of wine. Exercise 5.25 gives the least-squares line for predicting a nation’s heart disease death rate from its wine consumption. What is the predicted heart disease death rate for a country that drinks enough wine to supply 150 liters of alcohol per person? Explain why this result can’t be true. Explain why using the regression line for this prediction is not intelligent.

5.28 What’s my grade? In Professor Friedman’s economics course the correlation between the students’ total scores prior to the final examination and their final examination scores is $r = 0.6$. The pre-exam totals for all students in the course have mean 280 and standard deviation 30. The final exam scores have mean 75 and standard deviation 8. Professor Friedman has lost Julie’s final exam but knows that her total before the exam was 300. He decides to predict her final exam score from her pre-exam total.

(a) What is the slope of the least-squares regression line of final exam scores on pre-exam total scores in this course? What is the intercept?

- (b) Use the regression line to predict Julie's final exam score.
- (c) Julie doesn't think this method accurately predicts how well she did on the final exam. Use r^2 to argue that her actual score could have been much higher (or much lower) than the predicted value.

5.29 Going to class. A study of class attendance and grades among first-year students at a state university showed that in general students who attended a higher percent of their classes earned higher grades. Class attendance explained 16% of the variation in grade index among the students. What is the numerical value of the correlation between percent of classes attended and grade index?

5.30 Will I bomb the final? We expect that students who do well on the midterm exam in a course will usually also do well on the final exam. Gary Smith of Pomona College looked at the exam scores of all 346 students who took his statistics class over a 10-year period.¹² The least-squares line for predicting final exam score from midterm exam score was $\hat{y} = 46.6 + 0.41x$.

Octavio scores 10 points above the class mean on the midterm. How many points above the class mean do you predict that he will score on the final? (Hint: Use the fact that the least-squares line passes through the point $(\bar{x}, \bar{y})$ and the fact that Octavio's midterm score is $\bar{x} + 10$. This is an example of the phenomenon that gave "regression" its name: students who do well on the midterm will on the average do less well, but still above average, on the final.)

5.31 Height and reading score. A study of elementary school children, ages 6 to 11, finds a high positive correlation between height x and score y on a test of reading comprehension. What explains this correlation?

5.32 Do artificial sweeteners cause weight gain? People who use artificial sweeteners in place of sugar tend to be heavier than people who use sugar. Does this mean that artificial sweeteners cause weight gain? Give a more plausible explanation for this association.

5.33 What explains grade inflation? Students at almost all colleges and universities get higher grades than was the case 10 or 20 years ago. Is grade inflation caused by lower grading standards? Suggest some lurking variables that might affect the distribution of grades even if standards have remained the same.

5.34 The benefits of foreign language study. Members of a high school language club believe that study of a foreign language improves a student's command of English. From school records, they obtain the scores on an English achievement test given to all seniors. The mean score of seniors who studied a foreign language for at least two years is much higher than the mean score of seniors who studied no foreign language. These data are not good evidence that language study strengthens English skills. Identify the explanatory and response



(Daniel J. Cox/Natural Exposures)

variables in this study. Then explain what lurking variable prevents the conclusion that language study improves students' English scores.

5.35 Beware correlations based on averages. The variables used for regression and correlation are sometimes averages of a number of individual values. For example, both degree-days and gas consumption for the Sanchez household (Exercise 4.16) are averages over the days of a month. The values for individual days vary about the monthly average. If you calculated the correlation for the 485 days in these 16 months, would r be closer to 1 or closer to 0 than the r for the 16 monthly averages? Why?

5.36 Beavers and beetles. Ecologists sometimes find rather strange relationships in our environment. One study seems to show that beavers benefit beetles. The researchers laid out 23 circular plots, each 4 meters in diameter, in an area where beavers were cutting down cottonwood trees. In each plot, they counted the number of stumps from trees cut by beavers and the number of clusters of beetle larvae. Here are the data:¹³

Stumps	2	2	1	3	3	4	3	1	2	5	1	3
Beetle larvae	10	30	12	24	36	40	43	11	27	56	18	40
Stumps	2	1	2	2	1	1	4	1	2	1	4	
Beetle larvae	25	8	21	14	16	6	54	9	13	14	50	

- (a) Make a scatterplot that shows how the number of beaver-caused stumps influences the number of beetle larvae clusters. What does your plot show? (Ecologists think that the new sprouts from stumps are more tender than other cottonwood growth, so that beetles prefer them.)
- (b) Find the least-squares regression line and draw it on your plot.
- (c) What percent of the observed variation in beetle larvae counts can be explained by straight-line dependence on stump counts?

5.37 A computer game. A multimedia statistics learning system includes a test of skill in using the computer's mouse. The software displays a circle at a random location on the computer screen. The subject tries to click in the circle with the mouse as quickly as possible. A new circle appears as soon as the subject clicks the old one. Table 5.3 gives data for one subject's trials, 20 with each hand. Distance is the distance from the cursor location to the center of the new circle, in units whose actual size depends on the size of the screen. Time is the time required to click in the new circle, in milliseconds.¹⁴

- (a) We suspect that time depends on distance. Make a scatterplot of time against distance, using separate symbols for each hand.
- (b) Describe the pattern. How can you tell that the subject is right-handed?

TABLE 5.3 Reaction times in a computer game

Time	Distance	Hand	Time	Distance	Hand
115	190.70	right	240	190.70	left
96	138.52	right	190	138.52	left
110	165.08	right	170	165.08	left
100	126.19	right	125	126.19	left
111	163.19	right	315	163.19	left
101	305.66	right	240	305.66	left
111	176.15	right	141	176.15	left
106	162.78	right	210	162.78	left
96	147.87	right	200	147.87	left
96	271.46	right	401	271.46	left
95	40.25	right	320	40.25	left
96	24.76	right	113	24.76	left
96	104.80	right	176	104.80	left
106	136.80	right	211	136.80	left
100	308.60	right	238	308.60	left
113	279.80	right	316	279.80	left
123	125.51	right	176	125.51	left
111	329.80	right	173	329.80	left
95	51.66	right	210	51.66	left
108	201.95	right	170	201.95	left

- (c) Find the regression line of time on distance separately for each hand. Draw these lines on your plot. Which regression does a better job of predicting time from distance? Give numerical measures that describe the success of the two regressions.
- 5.38 Using residuals.** It is possible that the subject in Exercise 5.37 got better in later trials due to learning. It is also possible that he got worse due to fatigue. Plot the residuals from each regression against the time order of the trials (down the columns in Table 5.3). Is either of these systematic effects of time visible in the data?
- 5.39 How residuals behave.** Return to the merlin data regression of Exercise 5.22. Use your calculator or software to obtain the residuals. The residuals are the part of the response left over after the straight-line tie to the explanatory variable is removed. Find the correlation between the residuals and the explanatory variable. Your result should not be a surprise.
- 5.40 Using residuals.** Make a residual plot (residual against explanatory variable) for the merlin regression of Exercise 5.22. Use a y scale from -20 to 20 or wider to better see the pattern. Add a horizontal line at $y = 0$, the mean of the residuals.
- (a) Describe the pattern if we ignore the two years with $x = 38$. Do the $x = 38$ years fit this pattern?

- (b) Return to the original data. Make a scatterplot with two least-squares lines: with all nine years and without the two $x = 38$ years. Although the original regression in Exercise 5.22 seemed satisfactory, the two $x = 38$ years are influential. We would like more data for years with x greater than 33.

5.41 Using residuals. Return to the regression of highway mileage on city mileage in Exercise 5.3 (page 109). Use your calculator or software to obtain the residuals. Make a residual plot (residuals against city mileage) and add a horizontal line at $y = 0$ (the mean of the residuals).

- (a) Which car has the largest positive residual? The largest negative residual?
- (b) The Honda Insight, an extreme outlier, does not have the largest residual in either direction. Why is this not surprising?
- (c) Explain briefly what a large positive residual says about a car. What does a large negative residual say?

Chapter 5 MEDIA EXERCISES



5.42 Influence in regression. The *Correlation and Regression* applet allows you to create a scatterplot and to move points by dragging with the mouse. Click to create a group of 10 points in the lower-left corner of the scatterplot with a strong straight-line pattern (correlation about 0.9). Click the “Show least-squares line” box to display the regression line.

- (a) Add one point at the upper right that is far from the other 10 points but exactly on the regression line. Why does this outlier have no effect on the line even though it changes the correlation?
- (b) Now drag this last point down until it is opposite the group of 10 points. You see that one end of the least-squares line chases this single point, while the other end remains near the middle of the original group of 10. What makes the last point so influential?



5.43 Is regression useful? In Exercise 4.32 (page 102) you used the *Correlation and Regression* applet to create three scatterplots having correlation about $r = 0.7$ between the horizontal variable x and the vertical variable y . Create three similar scatterplots again, and click the “Show least-squares line” box to display the regression lines. Correlation $r = 0.7$ is considered reasonably strong in many areas of work. Because there is a reasonably strong correlation, we might use a regression line to predict y from x . In which of your three scatterplots does it make sense to use a straight line for prediction?



5.44 Guessing a regression line. Click on the scatterplot to create a group of 15 to 20 points from lower left to upper right with a clear positive straight-line pattern (correlation around 0.7). Click the “Draw line” button and use the mouse (right-click and drag) to draw a line through

the middle of the cloud of points from lower left to upper right. Note the “thermometer” above the plot. The red portion is the sum of the squared vertical distances from the points in the plot to the least-squares line. The green portion is the “extra” sum of squares for your line—it shows by how much your line misses the smallest possible sum of squares.

- (a) You drew a line by eye through the middle of the pattern. Yet the right-hand part of the bar is probably almost entirely green. What does that tell you?
- (b) Now click the “Show least-squares line” box. Is the slope of the least-squares line smaller (the new line is less steep) or larger (line is steeper) than that of your line? If you repeat this exercise several times, you will consistently get the same result. The least-squares line minimizes the *vertical* distances of the points from the line. It is *not* the line through the “middle” of the cloud of points. This is one reason why it is hard to draw a good regression line by eye.

5.45 An influenza epidemic. In 1918 and 1919 a worldwide outbreak of influenza killed more than 25 million people. The EESEE story “Influenza Outbreak of 1918” includes the following data on the number of new influenza cases and the number of deaths from the epidemic in San Francisco week by week from October 5, 1918, to January 25, 1919. The date given is the last day of the week.



Date	Oct. 5	Oct. 12	Oct. 19	Oct. 26	Nov. 2	Nov. 9	Nov. 16	Nov. 23	Nov. 30
Cases	36	531	4233	8682	7164	2229	600	164	57
Deaths	0	0	130	552	738	414	198	90	56

Date	Dec. 7	Dec. 14	Dec. 21	Dec. 28	Jan. 4	Jan. 11	Jan. 18	Jan. 25
Cases	722	1517	1828	1539	2416	3148	3465	1440
Deaths	50	71	137	178	194	290	310	149

- We expect the number of deaths to lag behind the number of new cases because the disease takes some time to kill its victims.
- (a) Make three scatterplots of deaths (the response variable) against each of new cases the same week, new cases one week earlier, and new cases two weeks earlier. Describe and compare the patterns you see.
 - (b) Find the correlations that go with your three plots.
 - (c) What do you conclude? Do the cases data predict deaths best with no lag, a one-week lag, or a two-week lag?
 - (d) Find the least-squares line for predicting weekly deaths for the choice of explanatory variable that gives the best predictions.

CHAPTER 6



(Digital Vision/PictureQuest)

In this chapter we cover...

Marginal distributions
Relationships between
categorical variables
Conditional distributions
Simpson's paradox

Two-Way Tables*

We have concentrated on relationships in which at least the response variable is quantitative. Now we will shift to describing relationships between two or more categorical variables. Some variables—such as gender, race, and occupation—are categorical by nature. Other categorical variables are created by grouping values of a quantitative variable into classes. Published data often appear in grouped form to save space. To analyze categorical data, we use the *counts* or *percents* of individuals that fall into various categories.

EXAMPLE 6.1 Age and education

Table 6.1 presents Census Bureau data for the year 2000 on the level of education reached by Americans of different ages.¹ Many people under 25 years of age have not completed their education, so they are left out of the table. Both variables, age and education, are grouped into categories. This is a **two-way table** because it describes two categorical variables. Education is the **row variable** because each row in the table describes people with one level of education. Age is the **column variable** because each column describes one age group. The entries in the table are the counts of persons in each age-by-education class. Although both age and education in this table are categorical variables, both have a natural order from least to most. The order of the rows and the columns in Table 6.1 reflects the order of the categories.

*This material is important in statistics, but it is needed later in this book only for Chapter 20. You may omit it if you do not plan to read Chapter 20 or delay reading it until you reach Chapter 20.

TABLE 6.1 Years of school completed, by age (thousands of persons)

Education	Age group			Total
	25 to 34	35 to 54	55 and over	
Did not complete high school	4,459	9,174	14,226	27,859
Completed high school	11,562	26,455	20,060	58,077
College, 1 to 3 years	10,693	22,647	11,125	44,465
College, 4 or more years	11,071	23,160	10,597	44,828
Total	37,786	81,435	56,008	175,230

Marginal distributions

How can we best grasp the information contained in Table 6.1? First, *look at the distribution of each variable separately*. The distribution of a categorical variable says how often each outcome occurred. The “Total” column at the right of the table contains the totals for each of the rows. These row totals give the distribution of education level (the row variable) among all people 25 years of age and older: 27,859,000 did not complete high school, 58,077,000 finished high school but did not attend college, and so on. In the same way, the “Total” row at the bottom of the table gives the age distribution. If the row and column totals are missing, the first thing to do in studying a two-way table is to calculate them. The distributions of education alone and age alone are called **marginal distributions** because they appear at the right and bottom margins of the two-way table.

marginal distribution

If you check the row and column totals in Table 6.1, you will notice some discrepancies. For example, the sum of the entries in the “25 to 34” column is 37,785. The entry in the “Total” row for that column is 37,786. The explanation is **roundoff error**. The table entries are in thousands of persons, and each is rounded to the nearest thousand. The Census Bureau obtained the “Total” entry by rounding the exact number of people aged 25 to 34 to the nearest thousand. The result was 37,786,000. Adding the row entries, each of which is already rounded, gives a slightly different result.

roundoff error

Percents are often more informative than counts. We can display the marginal distribution of education level in terms of percents by dividing each row total by the table total and converting to a percent.

EXAMPLE 6.2 Calculating a marginal distribution

The percent of people 25 years of age and older who have at least 4 years of college is

$$\frac{\text{total with 4 years of college}}{\text{table total}} = \frac{44,828}{175,230} = 0.256 = 25.6\%$$

136 CHAPTER 6 • Two-Way Tables

Do three more such calculations to obtain the marginal distribution of education level in percents. Here it is:

	Did not complete high school	Completed high school	1 to 3 years of college	≥ 4 years of college
Percent	15.9	33.1	25.4	25.6

The total is 100% because everyone is in one of the four education categories.

Each marginal distribution from a two-way table is a distribution for a single categorical variable. As we saw in Chapter 1 (page 8), we can use a bar graph or a pie chart to display such a distribution. Figure 6.1 is a bar graph of the distribution of years of schooling. We see that people with at least some college education make up about half of the 25 and over population.

In working with two-way tables, you must calculate lots of percents. Here's a tip to help decide what fraction gives the percent you want. Ask, "What group represents the total that I want a percent of?" The count for that group is the denominator of the fraction that leads to the percent. In Example 6.2, we wanted a percent "of people aged 25 years and over," so the count of people aged 25 and over (the table total) is the denominator.

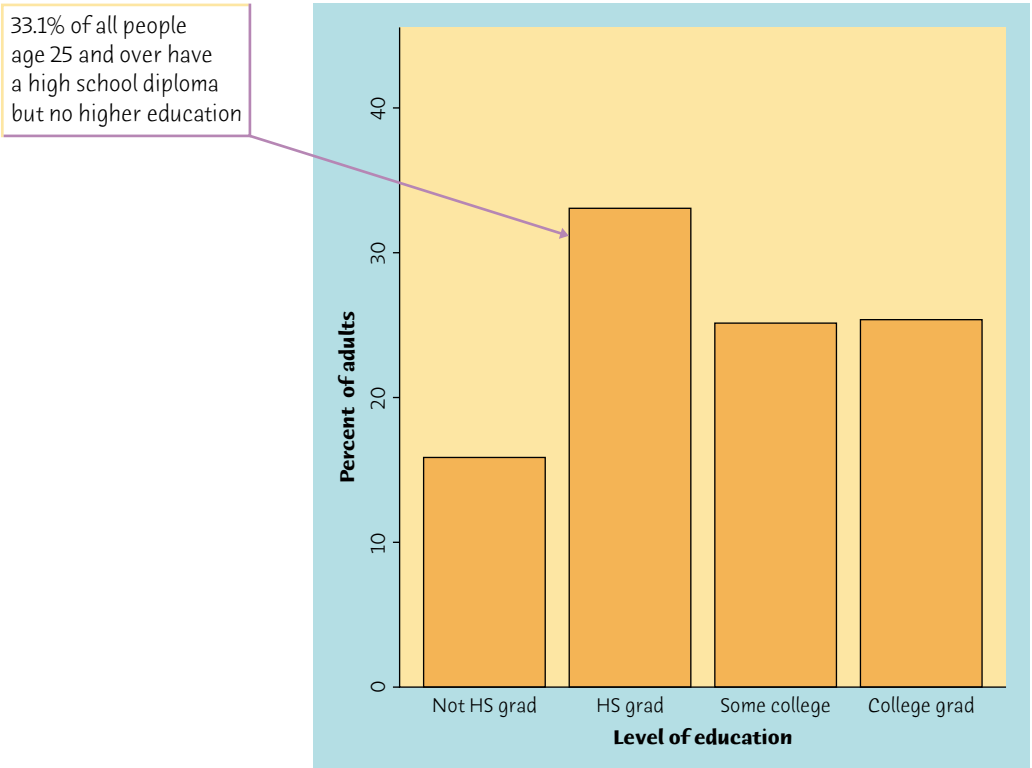


Figure 6.1 A bar graph of the distribution of years of schooling completed among people aged 25 years and over. This is one of the marginal distributions for Table 6.1.

APPLY YOUR KNOWLEDGE

- 6.1 **Age and education.** The counts in the “Total” column at the right of Table 6.1 are the counts of people in each education group. Explain why the sum of these counts is not exactly equal to 175,230, the table total that appears at the lower-right corner of the table.
- 6.2 **Age and education.** Give the marginal distribution of age (in percents) among people 25 years of age and older, starting from the counts in Table 6.1.
- 6.3 **Risks of playing soccer.** A study in Sweden looked at former elite soccer players, people who had played soccer but not at the elite level, and people of the same age who did not play soccer. Here is a two-way table that classifies these subjects by whether or not they had arthritis of the hip or knee by their mid-50s:²

	Elite	Non-elite	Did not play
Arthritis	10	9	24
No arthritis	61	206	548

- (a) How many people do these data describe?
- (b) How many of these people have arthritis of the hip or knee?
- (c) Give the marginal distribution of participation in soccer, both as counts and as percents.

Relationships between categorical variables

Table 6.1 contains much more information than the two marginal distributions of age alone and education alone. The nature of the relationship between age and education cannot be deduced from the separate distributions but requires the full table. **Relationships between categorical variables are described by calculating appropriate percents from the counts given.** We use percents because counts are often hard to compare. For example, 23,160,000 persons aged 35 to 54 have completed college, and only 10,597,000 persons in the 55-and-up age group have done so. These counts do not accurately describe the association, however, because there are many more people in the younger age group.

EXAMPLE 6.3 How common is college education?

What percent of people aged 25 to 34 have completed 4 years of college? This is the count of those who are 25 to 34 and have 4 years of college as a percent of the age group total:

$$\frac{11,071}{37,786} = 0.293 = 29.3\%$$

“People aged 25 to 34” is the group we want a percent of, so the count for that group is the denominator. In the same way, find the percent of people in each age group who have completed college. The comparison of all three groups is

	25 to 34	35 to 54	55 and over
Percent with 4 years of college	29.3	28.4	18.9

These percentages make it clear that a college education is less common among Americans over age 55 than among younger adults. This is an important aspect of the association between age and education.

These percents do *not* add to 100% because they are not the parts of a whole. They describe three different age groups.

APPLY YOUR KNOWLEDGE

- 6.4 Age and education.** Using the counts in Table 6.1, find the percent of people in each age group who did not complete high school. Draw a bar graph that compares these percents. State briefly what the data show.
- 6.5 Risks of playing soccer.** Find the percent of each group in the soccer-risk data of Exercise 6.3 who have arthritis. What do these percents say about the association between playing soccer and later arthritis?

Conditional distributions

Example 6.3 does not compare the complete distributions of years of schooling in the three age groups. It compares only the percents who finished college. Let’s look at the complete picture.

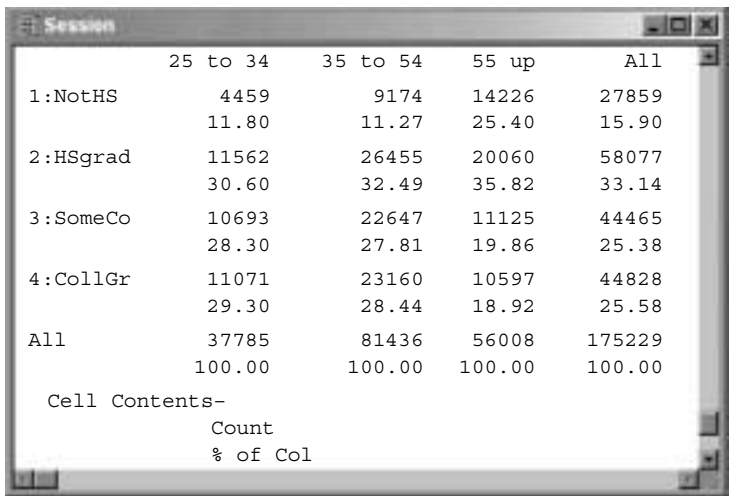
EXAMPLE 6.4 Calculating a conditional distribution

Information about the 25 to 34 age group occupies the first column in Table 6.1. To find the complete distribution of education in this age group, look only at that column. Compute each count as a percent of the column total, which is 37,786. Here is the distribution:

	Did not complete high school	Completed high school	1 to 3 years of college	≥4 years of college
Percent	11.8	30.6	28.3	29.3

conditional distribution

These percents do add to 100% because all 25- to 34-year-olds fall in one of the educational categories. The four percents together are the **conditional distribution** of education, given that a person is 25 to 34 years of age. We use the term “conditional” because the distribution refers only to people who satisfy the condition that they are 25 to 34 years old.



The image shows a Minitab Session window displaying a two-way table of education by age. The table has four columns representing age groups: '25 to 34', '35 to 54', '55 up', and 'All'. The rows represent education levels: '1:NotHS', '2:HSgrad', '3:SomeCo', '4:CollGr', and 'All'. Each cell in the table contains two values: a count and a percentage of the column total. The percentages in each column sum to 100%. At the bottom of the window, a legend indicates that the first value in each cell is the 'Count' and the second value is the '% of Col'.

	25 to 34	35 to 54	55 up	All
1:NotHS	4459 11.80	9174 11.27	14226 25.40	27859 15.90
2:HSgrad	11562 30.60	26455 32.49	20060 35.82	58077 33.14
3:SomeCo	10693 28.30	22647 27.81	11125 19.86	44465 25.38
4:CollGr	11071 29.30	23160 28.44	10597 18.92	44828 25.58
All	37785 100.00	81436 100.00	56008 100.00	175229 100.00

Cell Contents-
Count
% of Col

Figure 6.2 Minitab output of the two-way table of education by age with the three conditional distributions of education in each age group. The percents in each column add to 100%.

Now focus in turn on the second column (people aged 35 to 54) and then the third column (people 55 and over) of Table 6.1 in order to find two more conditional distributions. Statistical software can speed the task of finding each entry in a two-way table as a percent of its column total. Figure 6.2 displays the result. The software found the row and column totals from the table entries, so they differ slightly from those in Table 6.1. The software produces a legend at the bottom to remind us that the entries for each cell are the count and the count as a percent of its column total.

Each cell in this table contains a count from Table 6.1 along with that count as a percent of the column total. The percents in each column form the conditional distribution of years of schooling for one age group. The percents in each column add to 100% because everyone in the age group is accounted for. Comparing the conditional distributions reveals the nature of the association between age and education. The distributions of education in the two younger groups are quite similar, but higher education is less common in the 55-and-over group.

Bar graphs can help make the association visible. We could make three side-by-side bar graphs, each resembling Figure 6.1, to present the three conditional distributions. Figure 6.3 shows an alternative form of bar graph. Each set of three bars compares the percents in the three age groups who have reached a specific educational level. We see at once that the “25 to 34” and “35 to 54” bars are similar for all four levels of education, and that the “55 and over” bars show that many more people in this group did not finish high school and that many fewer have any college.

No single graph (such as a scatterplot) portrays the form of the relationship between categorical variables. No single numerical measure (such as the correlation) summarizes the strength of the association. Bar graphs are flexible enough to be helpful, but you must think about what comparisons you want

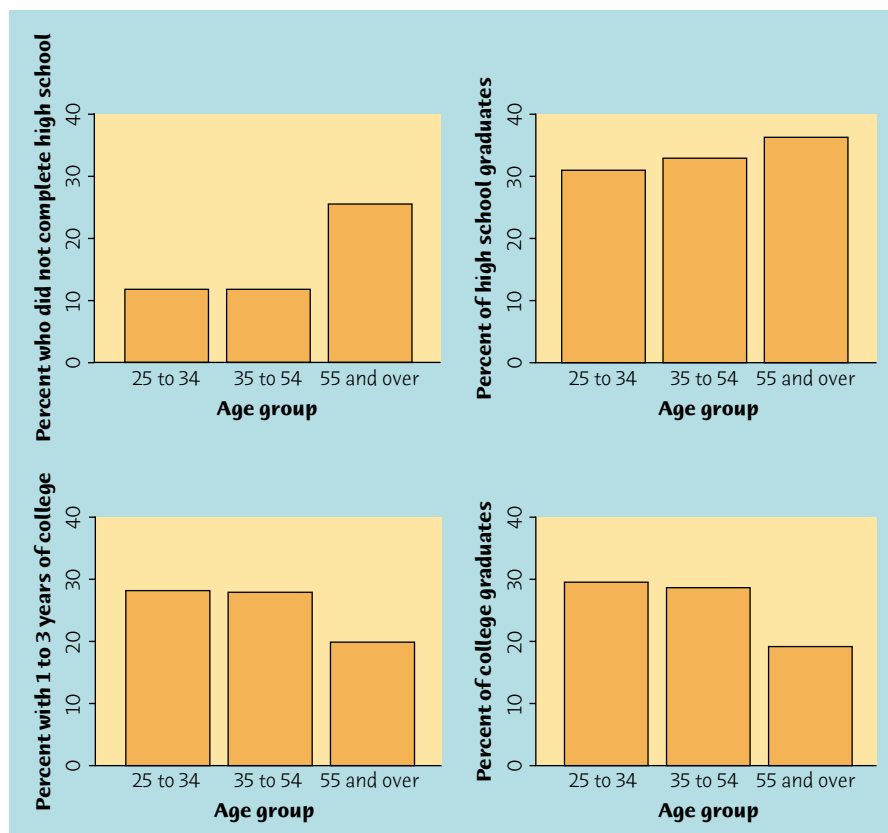


Figure 6.3 Bar graphs to compare the education levels of three age groups. Each graph compares the percents of three groups who fall in one of the four education levels.

to display. For numerical measures, we rely on well-chosen percents. You must decide which percents you need. Here is a hint: compare the conditional distributions of the response variable (education) for the separate values of the explanatory variable (age). That's what we did in Figure 6.3.

APPLY YOUR KNOWLEDGE

- 6.6 Education among older people.** Show the calculations to find the conditional distribution of education among persons 55 years old and over. Your results should agree with the “55 up” column in Figure 6.2 and with the heights of the “55 and over” bars in the four bar graphs in Figure 6.3.
- 6.7 Majors for men and women in business.** A study of the career plans of young women and men sent questionnaires to all 722 members of the senior class in the College of Business Administration at the University of Illinois. One question asked which major within the business program the student had chosen. Here are the data from the students who responded:³

	Female	Male
Accounting	68	56
Administration	91	40
Economics	5	6
Finance	61	59

- (a) Find the two conditional distributions of major, one for women and one for men. Based on your calculations, describe the differences between women and men with a graph and in words.
- (b) What percent of the students did not respond to the questionnaire? The nonresponse weakens conclusions drawn from these data.

6.8 Marginal distributions aren't the whole story. Here are the row and column totals for a two-way table with two rows and two columns:

a	b	50
c	d	50
60	40	100

Find *two different* sets of counts a , b , c , and d for the body of the table that give these same totals. This shows that the relationship between two variables cannot be obtained from the two individual distributions of the variables.

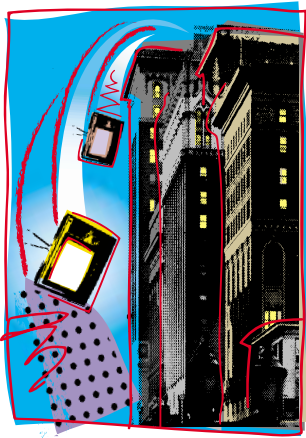
Simpson's paradox

As is the case with quantitative variables, the effects of lurking variables can change or even reverse relationships between two categorical variables. Here is an example that demonstrates the surprises that can await the unsuspecting user of data.

EXAMPLE 6.5 Do medical helicopters save lives?

Accident victims are sometimes taken by helicopter from the accident scene to a hospital. The helicopter may save time and also brings medical care to the accident scene. Does the use of helicopters save lives? We might compare the percents of accident victims who die with helicopter evacuation and with the usual transport to a hospital by road. Here are hypothetical data that illustrate a practical difficulty:⁴

	Helicopter	Road
Victim died	64	260
Victim survived	136	840
Total	200	1100



Attack of the killer TVs!

Despite occasional media frenzy, you won't find shark attacks high in any count of causes of death and injury. In fact, the 28 children killed by falling TV sets in the United States between 1990 and 1997 is four times the total number of people killed by great white sharks in the twentieth century.



(Ashley/Cooper/PICIMPACT/CORBIS)

We see that 32% (64 out of 200) helicopter patients died, compared with only 24% (260 out of 1100) of the others. That seems discouraging. The explanation is that the helicopter is sent mostly to serious accidents, so that the victims transported by helicopter are more often seriously injured than other victims. They are more likely to die with or without helicopter evacuation. Here are the same data broken down by the seriousness of the accident:

Serious Accidents			Less Serious Accidents		
	Helicopter	Road		Helicopter	Road
Died	48	60	Died	16	200
Survived	52	40	Survived	84	800
Total	100	100	Total	100	1000

Inspect these tables to convince yourself that they describe the same 1300 accidents as the original two-way table. For example, 200 were moved by helicopter, and 64 (48 + 16) of these died.

Among victims of serious accidents, the helicopter saves 52% (52 out of 100) compared with 40% for road transport. If we look only at less serious accidents, 84% of those transported by helicopter survive, versus 80% of those transported by road. Both groups of victims have a higher survival rate when evacuated by helicopter.

At first, it seems paradoxical that the helicopter does better for both groups of victims but worse when all victims are lumped together. Examining the data makes the explanation clear. Half the helicopter transport patients are from serious accidents, compared with only 100 of the 1100 road transport patients. So the helicopter carries patients who are more likely to die. The seriousness of the accident was a lurking variable that, until we uncovered it, made the relationship between survival and mode of transport to a hospital hard to interpret. Example 6.5 illustrates *Simpson’s paradox*.

SIMPSON’S PARADOX

An association or comparison that holds for all of several groups can reverse direction when the data are combined to form a single group. This reversal is called **Simpson’s paradox**.

The lurking variable in Simpson’s paradox is categorical. That is, it breaks the individuals into groups, as when accident victims are classified as injured in a “serious accident” or a “less serious accident.” Simpson’s paradox is just an extreme form of the fact that observed associations can be misleading when there are lurking variables.

APPLY YOUR KNOWLEDGE

6.9 Airline flight delays. Here are the numbers of flights on time and delayed for two airlines at five airports in one month. Overall on-time percentages for each airline are often reported in the news. The airport that flights serve is a lurking variable that can make such reports misleading.⁵

	Alaska Airlines		America West	
	On time	Delayed	On time	Delayed
Los Angeles	497	62	694	117
Phoenix	221	12	4840	415
San Diego	212	20	383	65
San Francisco	503	102	320	129
Seattle	1841	305	201	61

- (a) What percent of all Alaska Airlines flights were delayed? What percent of all America West flights were delayed? These are the numbers usually reported.
- (b) Now find the percent of delayed flights for Alaska Airlines at each of the five airports. Do the same for America West.
- (c) America West does worse at *every one* of the five airports, yet does better overall. That sounds impossible. Explain carefully, referring to the data, how this can happen. (The weather in Phoenix and Seattle lies behind this example of Simpson's paradox.)

6.10 Race and the death penalty. Whether a convicted murderer gets the death penalty seems to be influenced by the race of the victim. Here are data on 326 cases in which the defendant was convicted of murder:⁶

	White Defendant			Black Defendant	
	White victim	Black victim		White victim	Black victim
Death	19	0	Death	11	6
Not	132	9	Not	52	97

- (a) Use these data to make a two-way table of defendant's race (white or black) versus death penalty (yes or no).
- (b) Show that Simpson's paradox holds: a higher percent of white defendants are sentenced to death overall, but for both black and white victims a higher percent of black defendants are sentenced to death.

- (c) Use the data to explain why the paradox holds in language that a judge could understand.

Chapter 6 SUMMARY

A **two-way table** of counts organizes data about two categorical variables. Values of the **row variable** label the rows that run across the table, and values of the **column variable** label the columns that run down the table. Two-way tables are often used to summarize large amounts of information by grouping outcomes into categories.

The **row totals** and **column totals** in a two-way table give the **marginal distributions** of the two individual variables. It is clearer to present these distributions as percents of the table total. Marginal distributions tell us nothing about the relationship between the variables.

To find the **conditional distribution** of the row variable for one specific value of the column variable, look only at that one column in the table. Find each entry in the column as a percent of the column total.

There is a conditional distribution of the row variable for each column in the table. Comparing these conditional distributions is one way to describe the association between the row and the column variables. It is particularly useful when the column variable is the explanatory variable.

Bar graphs are a flexible means of presenting categorical data. There is no single best way to describe an association between two categorical variables.

A comparison between two variables that holds for each individual value of a third variable can be changed or even reversed when the data for all values of the third variable are combined. This is **Simpson’s paradox**. Simpson’s paradox is an example of the effect of lurking variables on an observed association.

Chapter 6 EXERCISES

Marital status and job level. We sometimes hear that getting married is good for your career. Table 6.2 presents data from one of the studies behind this generalization. To avoid gender effects, the investigators

TABLE 6.2 Marital status and job level

Job grade	Marital status				Total
	Single	Married	Divorced	Widowed	
1	58	874	15	8	955
2	222	3927	70	20	4239
3	50	2396	34	10	2490
4	7	533	7	4	551
Total	337	7730	126	42	8235

looked only at men. The data describe the marital status and the job level of all 8235 male managers and professionals employed by a large manufacturing firm.⁷ The firm assigns each position a grade that reflects the value of that particular job to the company. The authors of the study grouped the many job grades into quarters. Grade 1 contains jobs in the lowest quarter of the job grades, and Grade 4 contains those in the highest quarter. Exercises 6.11 to 6.15 are based on these data.

- 6.11 Marginal distributions.** Give (in percents) the two marginal distributions, for marital status and for job grade. Do each of your two sets of percents add to exactly 100%? If not, why not?
- 6.12 Percents.** What percent of single men hold Grade 1 jobs? What percent of Grade 1 jobs are held by single men?
- 6.13 Conditional distribution.** Give (in percents) the conditional distribution of job grade among single men. Should your percents add to 100% (up to roundoff error)?
- 6.14 Marital status and job grade.** One way to see the relationship is to look at who holds Grade 1 jobs.
- (a) There are 874 married men with Grade 1 jobs, and only 58 single men with such jobs. Explain why these counts by themselves don't describe the relationship between marital status and job grade.
 - (b) Find the percent of men in each marital status group who have Grade 1 jobs. Then find the percent in each marital group who have Grade 4 jobs. What do these percents say about the relationship?
- 6.15 Association is not causation.** The data in Table 6.2 show that single men are more likely to hold lower-grade jobs than are married men. We should not conclude that single men can help their career by getting married. What lurking variables might help explain the association between marital status and job grade?
- 6.16 Smoking by students and their parents.** Here are data from eight high schools on smoking among students and among their parents:⁸

	Neither parent smokes	One parent smokes	Both parents smoke
Student does not smoke	1168	1823	1380
Student smokes	188	416	400

- (a) How many students do these data describe?
- (b) What percent of these students smoke?
- (c) Calculate and compare percents to show how parents' smoking influences students' smoking. Briefly state your conclusions about the relationship.

6.17 Python eggs. How is the hatching of water python eggs influenced by the temperature of the snake’s nest? Researchers assigned newly laid eggs to one of three temperatures: hot, neutral, or cold. Hot duplicates the extra warmth provided by the mother python, and cold duplicates the absence of the mother. Here are the data on the number of eggs and the number that hatched:⁹

	Cold	Neutral	Hot
Number of eggs	27	56	104
Number hatched	16	38	75

- (a) Make a two-way table of temperature by outcome (hatched or not).
(b) Calculate the percent of eggs in each group that hatched. The researchers anticipated that eggs would not hatch at cold temperatures. Do the data support that anticipation?

6.18 Firearm deaths. Firearms are second to motor vehicles as a cause of nondisease deaths in the United States. Here are counts from a study of all firearm-related deaths in Milwaukee, Wisconsin, between 1990 and 1994.¹⁰ We want to compare the types of firearms used in homicides and in suicides. We suspect that long guns (shotguns and rifles) will more often be used in suicides because many people keep them at home for hunting. Make a careful comparison of homicides and suicides with a bar graph. What do you find about long guns versus handguns?

	Handgun	Shotgun	Rifle	Unknown	Total
Homicides	468	28	15	13	524
Suicides	124	22	24	5	175

6.19 Helping cocaine addicts. Cocaine addiction is hard to break. Addicts need cocaine to feel any pleasure, so perhaps giving them an antidepressant drug will help. A 3-year study with 72 chronic cocaine users compared an antidepressant drug called desipramine with lithium and a placebo. (Lithium is a standard drug to treat cocaine addiction. A placebo is a dummy drug, used so that the effect of being in the study but not taking any drug can be seen.) One-third of the subjects, chosen at random, received each drug. Here are the results:¹¹

	Desipramine	Lithium	Placebo
Relapse	10	18	20
No relapse	14	6	4
Total	24	24	24

- (a) Compare the effectiveness of the three treatments in preventing relapse. Use percents and draw a bar graph.
- (b) Do you think that this study gives good evidence that desipramine actually *causes* a reduction in relapses?

6.20 Female college professors. Purdue University is a Big Ten university that emphasizes engineering, scientific, and technical fields. University faculty start as assistant professors, then are promoted to associate professor and eventually to professor. Here is a two-way table that breaks down Purdue’s 1621 faculty members in the 1998–1999 academic year by gender and academic rank:

	Female	Male	Total
Assistant professors	126	213	339
Associate professors	149	411	560
Professors	60	662	722
Total	335	1286	1621

- (a) Describe the relationship between rank and gender by finding and commenting on several percents.
- (b) One possible explanation for the association might be discrimination (women find it harder to win promotion to higher ranks). Suggest other possible explanations.

6.21 Do angry people have more heart disease? People who get angry easily tend to have more heart disease. That’s the conclusion of a study that followed a random sample of 12,986 people from three locations for about four years. All subjects were free of heart disease at the beginning of the study. The subjects took the Spielberger Trait Anger Scale test, which measures how prone a person is to sudden anger. Here are data for the 8474 people in the sample who had normal blood pressure.¹² CHD stands for “coronary heart disease.” This includes people who had heart attacks and those who needed medical treatment for heart disease.



(Henryk Kaiser/eStock
Photography/PictureQuest)

	Low anger	Moderate anger	High anger	Total
CHD	53	110	27	190
No CHD	3057	4621	606	8284
Total	3110	4731	633	8474

Present evidence from this two-way table that backs up the study’s conclusion about the relationship between anger and heart disease.

6.22 Which hospital is safer? To help consumers make informed decisions about health care, the government releases data about patient outcomes in hospitals. You want to compare Hospital A and Hospital B, which serve your community. Here are data on all patients undergoing surgery in a recent time period. The data include the condition of the patient (“good” or “poor”) before the surgery. “Survived” means that the patient lived at least 6 weeks following surgery.

Good Condition			Poor Condition		
	Hospital A	Hospital B		Hospital A	Hospital B
Died	6	8	Died	57	8
Survived	594	592	Survived	1443	192
Total	600	600	Total	1500	200

- (a) Compare percents to show that Hospital A has a higher survival rate for both groups of patients.
- (b) Combine the data into a single two-way table of outcome (“survived” or “died”) by hospital (A or B). The local paper reports just these overall survival rates. Which hospital has the higher rate?
- (c) Explain from the data, in language that a reporter can understand, how Hospital B can do better overall even though Hospital A does better for both groups of patients.

6.23 Discrimination? Wabash Tech has two professional schools, business and law. Here are two-way tables of applicants to both schools, categorized by gender and admission decision. (Although these data are made up, similar situations occur in reality.)¹³

Business			Law		
	Admit	Deny		Admit	Deny
Male	480	120	Male	10	90
Female	180	20	Female	100	200

- (a) Make a two-way table of gender by admission decision for the two professional schools together by summing entries in these tables.
- (b) From the two-way table, calculate the percent of male applicants who are admitted and the percent of female applicants who are admitted. Wabash admits a higher percent of male applicants.
- (c) Now compute separately the percents of male and female applicants admitted by the business school and by the law school. Each school admits a higher percent of female applicants.
- (d) This is Simpson’s paradox: both schools admit a higher percent of the women who apply, but overall Wabash admits a lower percent of

female applicants than of male applicants. Explain carefully, as if speaking to a skeptical reporter, how it can happen that Wabash appears to favor males when each school individually favors females.

6.24 Obesity and health. Recent studies have shown that earlier reports underestimated the health risks associated with being overweight. The error was due to overlooking lurking variables. In particular, smoking tends both to reduce weight and to lead to earlier death. Illustrate Simpson’s paradox by a simplified version of this situation. That is, make up tables of overweight (yes or no) by early death (yes or no) by smoker (yes or no) such that

- Overweight smokers and overweight nonsmokers both tend to die earlier than those not overweight.
- But when smokers and nonsmokers are combined into a two-way table of overweight by early death, persons who are not overweight tend to die earlier.



(Aneal Vohra/Index Stock Imagery/PictureQuest)

Chapter 6 MEDIA EXERCISES

6.25 Baldness and heart attacks. The EESEE story “Baldness and Heart Attacks” reports results from comparing men under the age of 55 who survived a first heart attack with men admitted to the same hospitals for other reasons. As part of the study, each patient was asked to rate his own degree of baldness, from 1 (no baldness) to 5 (extreme baldness). Here are the counts:



Baldness score	Heart attack patients	Other patients
1	251	331
2	165	221
3	195	185
4	50	34
5	2	1

Do these data show a relationship between degree of baldness and having a heart attack? What kind of relationship? (The EESEE story points out that there are lurking variables—in particular, the ages of the men.)

